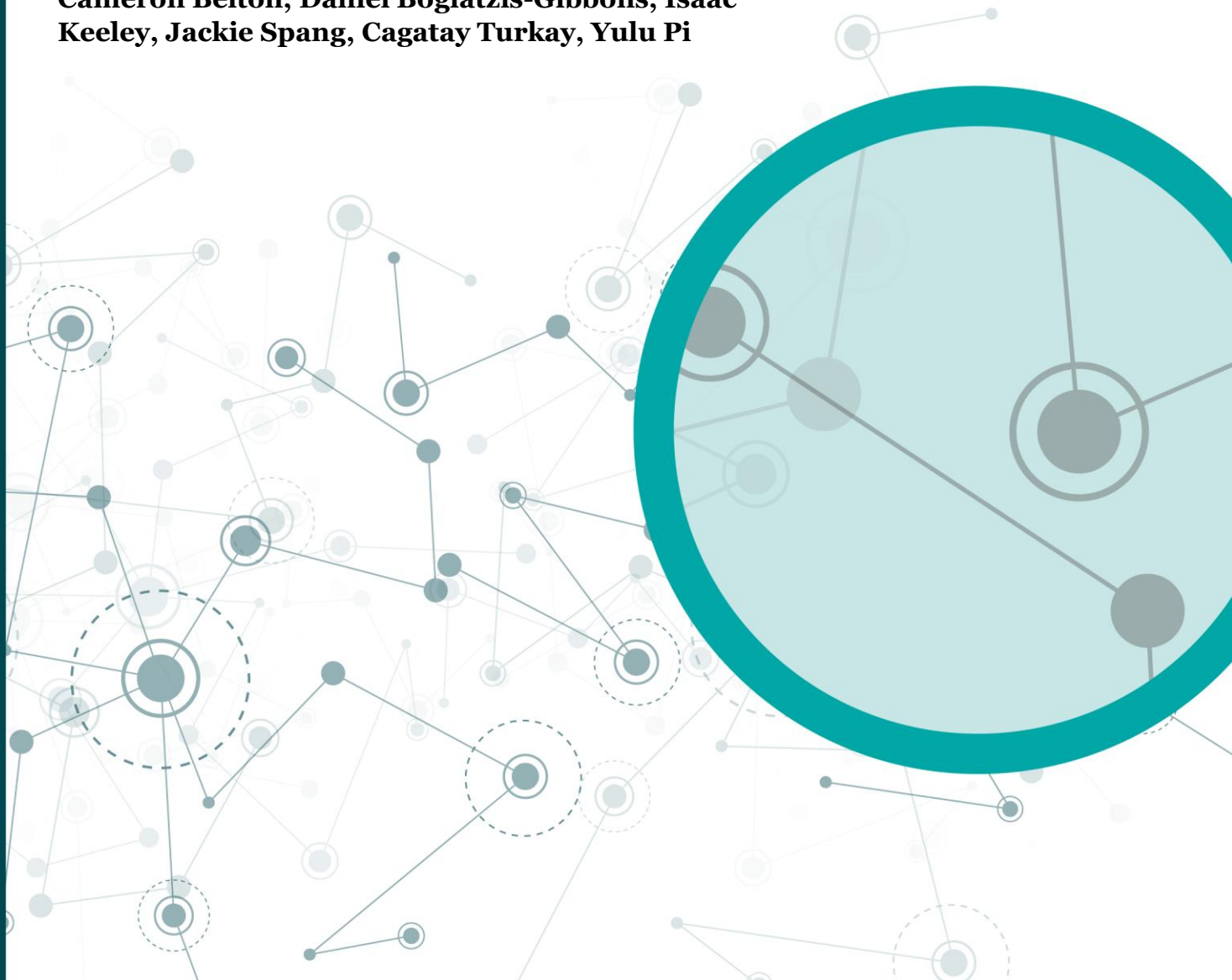
Financial Conduct Authority

Research Note

24 February 2025

Credit where credit is due: how can AI's role in credit decisions be explained?

Cameron Belton, Daniel Bogiatzis-Gibbons, Isaac Keeley, Jackie Spang, Cagatay Turkay, Yulu Pi



FCA research notes in financial regulation

The FCA research notes

The FCA is committed to encouraging debate on all aspects of financial regulation and to creating rigorous evidence to support its decision-making. To facilitate this, we publish a series of Research Notes, extending across economics and other disciplines.

The main factor in accepting papers is that they should make substantial contributions to knowledge and understanding of financial regulation. If you want to contribute to this series or comment on these papers, please contact David Stallibrass (david.stallibrass@fca.org.uk).

Disclaimer

Research notes contribute to the work of the FCA by providing rigorous research results and stimulating debate. While they may not necessarily represent the position of the FCA, they are one source of evidence that the FCA may use while discharging its functions and to inform its views. The FCA endeavours to ensure that research outputs are correct, through checks including independent referee reports, but the nature of such research and choice of research methods is a matter for the authors using their expert judgement. To the extent that research notes contain any errors or omissions, they should be attributed to the individual authors, rather than to the FCA.

Authors

Cameron, Isaac, Jackie, and Daniel were staff members of the FCA at time of writing. Daniel is in addition a PhD student in AI policy at Birkbeck College, London. Cagatay and Yulu are external academics from the University of Warwick.

Acknowledgements

We would like to thank the many internal reviewers of this document, as well as Ed Towers, Kieran Keohane, and Lawrence Charles for managerial support.

All our publications are available to download from www.fca.org.uk. If you would like to receive this paper in an alternative format, please call 020 7066 9644 or email publications_graphics@fca.org.uk or write to Editorial and Digital Department, Financial Conduct Authority, 12 Endeavour Square, London E20 1JN.

Contents

Summary

Artificial Intelligence (AI) offers the possibility of innovation and productivity across the financial services sector, yet also brings risks. This research note is part of the FCA's AI Research Series, a program of publications designed to take the conversation around AI forward.

The note explores the issue of AI explainability in the context of algorithm-assisted decision-making, using consumer credit decisions as a case study to test out different approaches. We used an online experiment to study whether different kinds, or 'genres', of explanation lead to better consumer outcomes such as consumers' ability to judge whether algorithm-assisted decisions are erroneous. Specifically, we tested whether participants were able to identify errors caused either by incorrect data used by the algorithm or by flaws in the algorithm's decision logic itself.

We tested four explanation genres:

- a data-centric explanation, which provided an overview of all the data available to the algorithm
- a features-based explanation, which explained only which variables or features were important to decision-making
- a combination of aspects of the data-centric and features-based explanations
- a combination of aspects of the data-centric and features-based explanations with the addition of the exact rule used to determine an individual's creditworthiness

The method of explaining algorithm-assisted decisions significantly impacted participants' ability to judge these decisions. On average, the data-centric explanation was the most effective genre. However, we found that the impact of our explanation genres varied depending on whether there was an error in the algorithm's decision-making and the type of error. For example, while the data-centric explanation helped participants challenge errors in the algorithm's decision logic, such as the algorithm failing to use a relevant piece of information about the consumer, we found that it impaired participants' ability to identify incorrect data input.

We propose two possible hypotheses to explain the inconsistent effects of our explanation genres in this study: the salience of errors and the role of confirmatory information. The salience of errors hypothesis suggests that additional information may make it more difficult to spot errors because there is simply more information to review. The confirmatory information hypothesis suggests that additional information about the algorithm's decision logic may encourage participants to focus on whether this decision logic was followed rather than if the decision logic was sound. In both cases, the hypothesis suggests that additional information may harm consumers' ability to judge algorithm-assisted decisions.

This experiment also found that participants who were given more information about the inner workings of the algorithm's decision-making reported feeling more confident in their ability to judge the algorithm's decisions. However, their actual judgement was

worse on average. This highlights an important potential disconnect between objective performance and participants' self-reported abilities.

At face value, providing additional information about the inner workings of the algorithm may be well received by consumers. However, the research finds that more information may not always be helpful for decision-making and could lead to worse outcomes for consumers by impairing their ability to challenge errors. When and where this is true clearly depends on the specific context.

More generally, we acknowledge that the findings and insights from our research may be specific to the context and design of our experiment, and that the effectiveness of any approach to explainability is likely to depend on the particular circumstances.

However, taken together, our findings reiterate the value of testing accompanying materials that may be provided to consumers when explaining AI, ML and/or algorithmic decision-making to understand how effective they are. Our findings also underscore the importance of testing consumers' decision-making within the relevant context, rather than relying solely on self-reported attitudes.

To advance the conversation on AI explainability we welcome further research to explore approaches to explaining AI assisted decisions in other contexts within financial services, the specific mechanisms for how explainability methods may impact consumers, alternative ways of presenting explanation genres, and the broader consumer journey beyond recognising errors.

Glossary

AI: AI stands for 'Artificial Intelligence' and refers to the development of computer systems that mimic or attempt to surpass human intellectual abilities. There is not an established consensus on what technologies constitute AI, but it could include ChatGPT, facial recognition software, and predictive systems like credit scoring.

Arrears: In the context of consumer credit, where a customer has not made a payment after the due date has passed, which could eventually lead to the account being passed onto a debt collector or a lender taking action to enforce the loan agreement.

Credit Scoring: A modelling process that produces a numerical expression, or binary decision, for a customer's creditworthiness, that is if they're at a low enough risk of going into arrears or defaulting on credit to be worth lending to by a credit firm.

Algorithm: A precise finite set of rules to follow to achieve some outcome, for example simple addition is an algorithm. In an AI context, it refers to the output of the modelling process which could be used to make or assist decisions about consumers.

Decision Tree: A simple form of predictive model where a small number of variables and decision rules (for example, income over £50,000) are used to divide consumers into different predicted probabilities of default or arrears.

Explainability: A topic within the study of AI concerned with what elements of a complex algorithmic decision can be retrieved for a consumer or a firm to understand.

Supervised Machine Learning: The discipline of sophisticated mathematical techniques that create algorithms for predicting some future outcome about a consumer (or more generally, firm or other unit) based on training data on past consumers.

1 Introduction and policy context

Artificial Intelligence (AI), including Supervised Machine Learning (SML) can be used to generate predictions of outcomes, including the probability of default. While these systems promise improved decision accuracy and financial innovation, there are concerns around bias, explainability, transparency and accountability.

We are publishing a series of FCA Research Notes on AI to spark discussion on these issues, drawing on a variety of different regulatory and academic perspectives. Research notes contribute to FCA objectives by providing rigorous research results and stimulating debate. They represent the views of the authors, not the FCA, and to the extent that research notes contain any errors or omissions, they should be attributed to the individual authors.

Transparency and explainability in AI systems

This research note focuses on the issue of transparency in AI and ML systems and its potential impact on consumer outcomes. These systems are often viewed as 'black boxes' because their internal workings and decision-making processes are hard for users to interpret. This lack of clarity makes it difficult for consumers to understand how they operate (see Lipton, 2018 for a discussion), which has led to calls for greater transparency and explainability of their predictions (see Morley et al., 2019; Guidotti et al., 2018).

Despite becoming increasingly sophisticated, AI and ML systems can be susceptible to mistakes and in contexts where the decisions made by these systems have financial implications, these mistakes could be costly and have adverse impacts on consumers (see Fuster et al., 2021; Barocas et al., 2021).

One proposed solution to the lack of transparency in AI and ML systems is AI explainability (also known as 'explainable AI' or 'XAI'), which refers to the concept of making AI systems understandable. Explainability seeks to provide people with clear reasons or justifications for AI's decisions, to "enable end users to understand, appropriately trust, and effectively manage the emerging generation of AI systems" (Gunning et al., 2021).

Explainability involves various techniques, ranging from simply explaining the data used to make a decision, to highlighting the factors that influenced the specific decision (Ribeiro et al., 2016), to providing explanations of how an entire model works, such as through permutation feature importance (Lundberg and Lee, 2017). Our focus is on supervised machine learning models and the effectiveness of different explainability methods in that context in promoting consumer comprehension of how models make mistakes.

There has been a lack of research conducted on the use of explainability techniques, especially within the financial services context. Little is known about how different explainability techniques can help consumers better understand model predictions or contest incorrect decisions. However, notably, Binns et al., (2018) show that the approach to explaining algorithmic decision-making matters less to perceptions of justice than the scenario in which machine learning is used. Additionally, Dodge et al., (2018) demonstrate in a small-scale experiment with 160 participants that fairness perceptions are influenced by the type of explanation provided.

The most similar work to the research presented in this research paper is Poursabzi-Sangdeh et al. (2021), who find that transparency in machine learning models may not always be helpful. In particular, they find that increasing the sheer amount of information presented to participants can make it harder for them to detect mistakes in a model.

Approach to our research

In this research, we combined various explainability techniques to develop four different approaches for explaining algorithmic decisions. For the purpose of this research, we refer to these approaches as 'explanation genres'. We tested these explanation genres in the context of creditworthiness, where the algorithmic decisions represented a simple AI system used to determine access to credit. We selected the context of creditworthiness as an example to test the effectiveness of different explainability approaches rather than to draw conclusions about explainability in this context specifically (also noting that the design of the experiment and the scenarios tested may not reflect how firms make credit lending decisions or present these to consumers).

The aim of the research was to examine, in principle, whether different explanation genres affected participants' ability to identify errors in algorithmic decision-making, promoted comprehension, and improved confidence in challenging errors. It aimed to contribute to the literature by assessing how our explanation genres impacted consumers ability to evaluate and understand algorithm-assisted decisions about their creditworthiness.

We have set out some of the interesting insights and findings from our research in Sections 3 and 4, acknowledging that these may be specific to the particular context and design of our experiment. Furthermore, the views and findings set out in this note are not intended to set any regulatory expectations or guidance about what firms or practitioners should do, how they should approach AI explainability or manage AI risks more generally. In all cases, firms will need to consider the risks relating to AI adoption in the context of their specific use cases and in light of applicable requirements.

In particular, as noted in the FCA's AI Update (2024), the FCA's existing regulatory framework does not speak directly to explainability of AI systems. However, there are a number of high-level requirements and principles relating to consumer protection that are relevant to the information firms provide to consumers. In particular, rules under the Consumer Duty on consumer understanding refer to meeting the information needs of retail customers and equipping them to make decisions that are effective, timely and properly informed. Those rules also require firms, where appropriate, to test and monitor the impact of communications to consumers, to identify whether they are supporting good outcomes. This research seeks to contribute to the limited body of evidence on explainability in AI and ML systems and to provide insights for those considering using explainability methods to communicate with their consumers.

The remainder of the paper contains the following sections: 2) Methodology, which outlines how the experiment worked, the outcomes we measured, the explanation genres we tested, and the representative sample we recruited; 3) Results, where we report our findings; and 4) Discussion, where we present some hypotheses to explain our findings and provide suggestions for further research.

2 Methodology

This section details the methodology we used to test our explanation genres. This includes the experimental flow, the explanation genres we tested, the hypothetical credit application scenarios we used, and our analytical strategy.

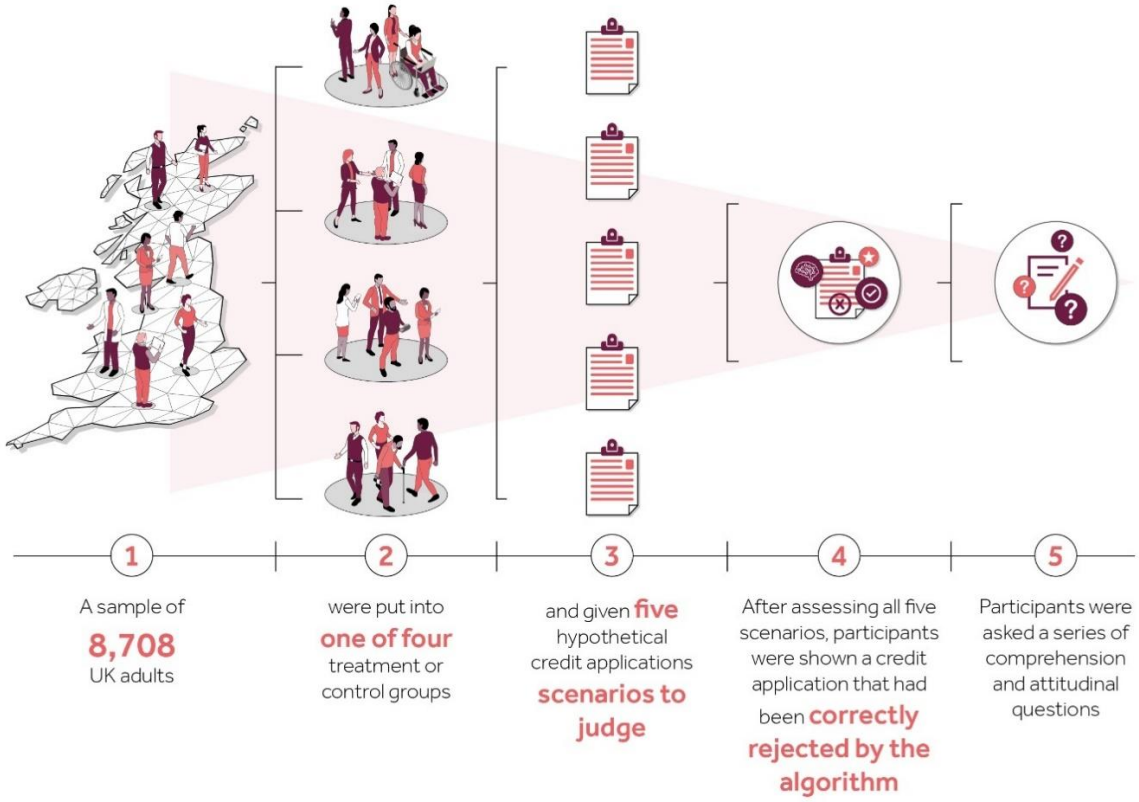
Experimental design

We recruited a sample of 8,860 UK adults through Prolific, an online panel provider. We conducted the experiment using Qualtrics, an online survey platform. We outline the experimental flow below (Figure 1) showing a high-level overview of participants' journey through the experiment.

Our experiment consisted of 3 parts: i) a judgement task, measuring participants' ability to judge whether algorithm-assisted creditworthiness decisions were correct for a hypothetical applicant; ii) comprehension questions, measuring participants' understanding of the algorithm's decision-making and iii) attitudinal questions, surveying participants' attitudes towards the task and explanation genres.

The structure of the experiment was the same for all participants. However, we gave participants different explanation genres to support them with the judgement task and comprehension questions. Participants were randomly assigned to one of 4 explanation genres prior to the task. This design meant that any differences between groups in performance on the task, or responses to the comprehension/attitudinal questions, could be attributed to the explanation genre provided.

Figure 1. Experimental flow



Judgement Task

First, participants completed the judgement task. Participants were presented with a series of hypothetical credit applications, which we refer to as 'scenarios'. Each scenario included:

- the profile of the hypothetical applicant (including the data inputs relating to the applicant that could be considered by the algorithm, which we refer to as 'features')
- the outcome of the application
- an explanation genre

The 5 profiles and the outcomes of their applications were the same across all participants. However, we gave participants different explanation genres to help them with the task, depending on the treatment or control group they were randomly assigned to. The explanation genres we tested are outlined in the Explanation Genres section below.

For each scenario, we asked participants to imagine themselves as the individual applying for credit and to act as though the information provided in the applicant's profile was correct. Participants were told that the credit provider had used an algorithm to help decide whether to lend and were warned that the algorithm could make errors (see Scenario section for more detail on the errors). The participants' task was to accept or challenge the algorithm's decision, based on their judgement of whether the decision was correct.

As we explain in the Scenario section, we created these scenarios and labelled decisions as 'correct' or 'incorrect' using a simplified credit default algorithm based on FCA credit

file data. Decisions were considered 'correct' or 'incorrect' depending on whether they matched our predictions of delinquency. Importantly, 'incorrect' decisions could happen either because of errors in how the algorithm was used (e.g., wrong data input) or because the algorithm's decision logic led to credit decisions that differed from our predictions, such that there was deemed an error in the decision logic itself. We acknowledge that by using this simplified credit algorithm our scenarios may not accurately reflect the approach of credit providers. Credit providers establish their own lending criteria where risk of delinquency is likely just one factor in their decisions.

Having judged the decision in each scenario, participants were also asked to select from a pre-populated list of reasons why they had accepted or challenged the decision (see Annex). Each participant repeated this process for the 5 different scenarios. The order of scenarios was randomised to mitigate the potential effects from ordering, learning, or experimental fatigue. Participants earned money for every correct answer they gave, both on the comprehension questions and judgement task, to incentivise their attention to the experiment.

Comprehension and Attitudinal Questions

After judging all 5 scenarios, participants were shown a credit application that had been correctly rejected by the algorithm. They were informed that this scenario contained no errors and should be used as the basis for answering the subsequent comprehension questions. Participants then answered 3 comprehension questions, detailed in the Annex.

After answering the comprehension questions, participants completed attitudinal questions which surveyed whether they found the information provided during the task helpful, sufficient, and important. Participants were also asked to indicate how confident they were in their ability to challenge credit decisions they thought were incorrect.

Explanation Genres

We tested 4 explanation genres. Table 1 provides a high-level summary of these explanation genres.

Table 1. Overview of explanation genres

Treatment	Summary
Data-centric explanation (control)	Described all the data inputs available to be used by the algorithm, including those not actually considered by the algorithm. Described the source of the data inputs and compared the individual's profile data to the average of past applicants.
Features-based explanation	Displayed only the features considered by the algorithm; highlighted the importance of each feature in the decision-making process; and indicated how each one influenced the likelihood of approval.
Combination: data-centric + features-based explanation	Showed the key data categories for features considered by the algorithm, their importance and direction in influencing the decision, whilst also providing details on the data distribution and the sources of the information used.
Combination + decision-rule explanation	The same as the combination of features-based and data-centric approach with the addition of a decision-rule (for example 'if X is greater than Y, then accept') to show how specific data features influence the likelihood of the application being approved. As the only difference between this treatment and the previous one is the inclusion of the decision-rule, comparing these two treatments allows us to isolate the impact of the decision-rule specifically.

As mentioned, there are no specific requirements or FCA expectations for firms to provide explanations of how AI or ML is used to determine creditworthiness. We could have used a no information control group from which to compare all explanation genres. However, instructing participants to judge the accuracy of credit decisions about hypothetical consumers without providing any information about how the decision was made would not have been particularly informative. We therefore chose to use the data-centric explanation as our control. The data-centric explanation is the only genre which does not explain how the algorithm considers each feature and was therefore the most logical choice as a baseline to compare against.

All our explanation genres:

- used simplified language
- adopted a standard presentation style, i.e. using tables to list the features
- included the same profile information
- had the same decision outcomes
- listed features in the same order

We maintained consistency in these elements to isolate the explanation genres' effects without interference from other factors, such as the ability to understand technical language, ability to interpret graphs/tables, or information ordering.

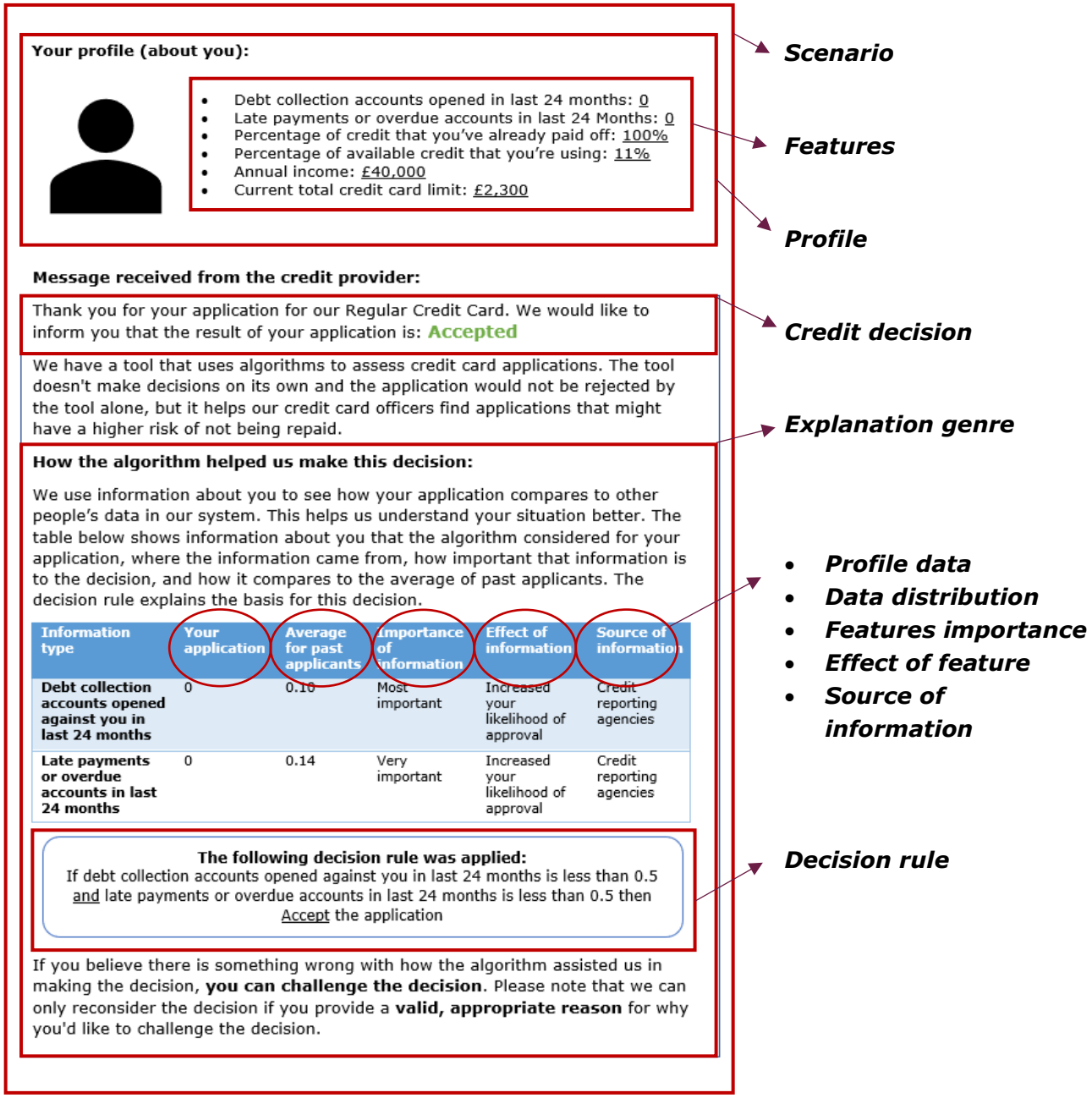
Explanation genres varied in the volume and complexity of information provided. The differences between genres are summarised in **Error! Reference source not found.**.

Table 2. Summary of differences between explanation genres

Characteristic	Description	Treatment Groups			
		Data-centric	Features-based	Combination	Combination + rule
Features included	The data inputs listed in the explanation.	All features	Only those considered by model	Only those considered by model	Only those considered by model
Feature importance	Displays relative importance of each feature in determining the outcome	No	Yes	Yes	Yes
Effect of feature	Displays whether each feature affects the likelihood of approval	No	Yes	Yes	Yes
Data distribution	Displays the average of past applicants' data for each data input	Yes	No	Yes	Yes
Source of information	Displays the source of each data input (e.g., credit agency, applicant)	Yes	No	Yes	Yes
Decision-rule	Value thresholds and logical statements that explain how the rule is governed	No	No	No	Yes

An example of the 'combination + decision-rule' explanation is shown below (Figure 2). Examples of all explanation genres are included in the Annex.

Figure 2. Scenario 1 for those shown the combination + rule explanation



Scenario design

To make the decisions in our scenarios realistic, we built a simplified credit default algorithm using FCA credit file data to predict the likelihood of a delinquency event (failure to make a scheduled payment) within the next 12 months. The supervised machine learning (ML) algorithm took the form of a decision tree and used several variables (detailed in the Annex) such as self-reported annual income and total credit card limit, as reflected in the explanation genres presented to participants.

Based on this algorithm and its decision tree, we created hypothetical scenarios in three categories: one scenario where the algorithm correctly predicted no delinquency event (i.e. a correct acceptance), one scenario where the algorithm correctly predicted delinquency events (i.e. a correct rejection), and 3 scenarios where the algorithm incorrectly predicted delinquency events (i.e. incorrect rejections).

For these 'incorrect rejections', we reviewed literature on common types of mistakes made by machine learning models and constructed these scenarios to reflect instances where the algorithm would likely mispredict a delinquency event due to those types of error. A table summarising these scenarios, including the credit decision associated with each scenario and whether they were correct or incorrect, is included below (Table 3). Examples of what these scenarios looked like can be found in the Annex.

These scenarios were designed using a simplified credit algorithm solely for the purpose of testing our explanation genres in an experimental setting. The algorithm and its predictions are intended to reflect simplified predictions of the actual risks of delinquency events and are not intended to reflect how AI creditworthiness decisions are made or should be presented to consumers in the consumer credit lending sector.

Table 3. Summary of scenarios

Scenario	Decision	Correct/Incorrect	Error type
1	Acceptance	Correct	N/A
2	Rejection	Correct	N/A
3	Rejection	Incorrect	Incorrect data input
4	Rejection	Incorrect	Overreliance on one feature
5	Rejection	Incorrect	Failure to consider relevant feature

In Scenario 3, the algorithm incorrectly rejected the credit application due to a data input error. Specifically, the data considered by the algorithm did not match the data in the hypothetical applicant's profile, leading the algorithm to attribute a higher risk of delinquency than should be the case. According to the explanations shown to participants, the data input into the algorithm suggested that the applicant had previously had a debt collection account opened against them. The applicant's profile showed that this was not the case. Notably, this was the only scenario where the participant could directly check that an objective error had occurred.

In scenarios 4 and 5, participants were required to identify errors in the algorithm itself (which they could not see), not an error in how the algorithm had been implemented. For these scenarios, participants therefore had to rely on their own judgement of whether the

credit decision was appropriate given the information in the applicants' profile and any of the information shown for their respective treatment group.

In Scenario 4, the algorithm incorrectly rejected the credit application because of an overreliance on one feature. Referring to the algorithm's decision tree, the algorithm only considered one feature (debt collection accounts opened against the applicant in the last 24 months) and did not consider the rest of the data in the case of this applicant's profile. This meant that it ignored favourable data such as the applicant's annual income (£180,000) or the percentage of credit being used (11%), which led to a misprediction of delinquency. Unlike identifying incorrect data input, participants could not directly check that an error had occurred. Rather than a clear inconsistency between the applicant's data and the data considered by the algorithm, the algorithm's overreliance on one feature was an error in the algorithm's decision logic. For example, for those given the decision-rule (a clear statement of the algorithm's decision logic), participants would have to challenge the decision logic itself to correctly challenge this error. They could not simply check whether the decision rule was followed properly.

Scenario 5 was incorrect due to the algorithm's failure to consider relevant features. In this case, despite the data being available to be used by the algorithm, the structure of the decision tree meant that the algorithm did not ultimately consider important features such as the applicant's annual income (£280,000), leading to a misprediction of that individual's delinquency. As with Scenario 4, participants could not directly check whether an error had occurred by simply applying the decision logic. Identifying that the algorithm failed to consider relevant features relied on the participant's own judgement of the algorithm's decision-making process and the individual's creditworthiness.

Outcome measures

Table 4 below details the specific outcome measures we examined, including a brief description of each outcome and the econometric method used to assess changes in those outcomes. We documented these, along with our empirical strategy, in our internal trial protocol prior to launching the experiment. Outcomes are classified as (1) Primary, (2) Secondary, or (3) Exploratory based on their role in the experiment: the Primary outcome was our main focus, Secondary outcomes provided broader contextual insight, and Exploratory outcomes helped understand differences in Primary and Secondary outcomes across our explanation genres.

Table 4. Outcome measures

Outcome	Description	Model Used	Classification/ Analysis Type
Performance on judgement task			
Correct judgements of whether algorithm-assisted decisions were correct	Score out of 5 of credit decisions judged correctly (all scenarios are weighted equally)	Ordinary Least Squares (OLS)	Primary

Likelihood of correct judgement across each scenario individually	Considered correct when participant correctly challenged/accepted decision	OLS	Secondary
Comprehension of algorithm's decision making			
Comprehension of basic information about how the algorithm is used (CQ1)	Considered correct when participant selected answer: "The algorithm compares the applicant's profile with similar profiles and flags any high risk profiles for manual review"	OLS	Secondary
Comprehension of directionality of 'debt collection accounts opened against you' feature information (CQ2)	Considered correct when participant selected answer: "Fewer debt collection accounts opened against you in the last 24 months increases the likelihood of approval"	OLS	Secondary
Comprehension of features importance information (CQ3)	Considered correct when participant selected "The number of debt collection accounts opened against you in the last 24 months"	OLS	Secondary
Attitudes toward the task and information provided			
Importance of information	Considered important when participant selected 'Slightly important' or 'Very important'	OLS	Exploratory
Helpfulness of information	Considered helpful when participant selected 'Slightly helpful' or 'Very helpful'	OLS	Exploratory
Sufficiency of information	Considered sufficient when participant selected 'Somewhat agree' or 'Strongly agree'	OLS	Exploratory
Confidence in ability to disagree with decision	Considered confident when participant selected 'Somewhat confident' or 'Very confident'	OLS	Exploratory

Empirical Strategy

We conducted an online randomised controlled trial (RCT) using a between-subject design. This meant that participants were randomly allocated to either the control group or one of our 3 treatment groups. This design allowed us to directly compare the effects of different explanation genres against the control (data-centric explanation) on the consumer outcomes measured.

The regression models employed in our analysis are detailed in Table 4. Outcome measures, with full model specifications are provided in the Annex. These models include covariates for age and sex assignment at birth. These models allowed us to test the relationship between the explanation genres and our outcome variables. By controlling for demographic factors, we were able to isolate the effects of our explanation genres, improving the robustness of our findings.

We corrected for multiple hypotheses testing using the Bonferroni correction approach (Abdi, 2007). This involved dividing the traditional significance threshold ($\alpha = 0.05$) by the number of comparisons made, which in this case was our 3 treatments compared to the control. Further details on our approach to multiple comparisons can be found in the Annex.

The design of our combination and 'combination + decision-rules' treatments allowed us to isolate the effect of adding a decision-rule by directly comparing outcome measures between these two treatments. This was the only other between-treatment comparison we were interested in, and we did not correct for multiple hypothesis testing for this comparison. As with the rest of our empirical strategy, we outlined this in our internal trial protocol prior to launching the experiment. We also ran a series of robustness checks and sensitivity analyses, detailed in the Annex.

Sample description and attrition

In our study, we collected responses from 8,860 UK adults, recruited via Prolific. We determined our target sample through power analysis, detailed in the Annex. The composition of our sample is described below.

- The 'sex assignment at birth' distribution was balanced, with a 50/50 split between male and female participants.
- The median age of participants was 38 years, closely matching the UK's median age of 40.6.
- Approximately 16% of participants identified as belonging to an ethnic minority background, which is broadly comparable to the 18% of the UK population.
- 50% of participants were in full-time employment, which is lower than the UK's overall employment rate of 75%.

We found that attrition, those dropping out of the experiment after starting it, was balanced across our treatment groups. Our overall attrition rate was low, with around 1.6% (N=141) dropping out. For the results we report below, we included those who dropped out of the experiment if they had been exposed to treatment, coding missing responses as 'wrong' answers. We also ran sensitivity analyses around this approach, such as only analysing complete cases. We found no noteworthy differences to those we

report below. More details on our approach to dealing with missing data can be found in the Annex.

3 Results

In this section, we share our findings on the 3 outcomes: accuracy of judgement, comprehension, and attitudes.

We found that, across all 5 scenarios, the control genre was the most effective for helping participants judge the scenarios accurately. However, the effect of our explanation genres on judgement depended on whether there was an error and what kind of error it was. The treatment genres helped participants detect incorrect data input but worsened their ability to detect the algorithm's overreliance on, or failure to consider, relevant features.

Participants' attitudes did not match this overall trend in performance. Despite being better overall at spotting errors, those in the control genre were less likely to report that they were confident in their ability to challenge incorrect decisions, or think the information provided was sufficient. In general, however, a large majority across all groups said that the information provided was important and helpful. Performance in objective comprehension questions across groups was varied. Which group performed best in these depended on the specific question asked.

Table 5 summarises how each explanation genre affected the outcomes measured. We compare each genre to the control group, who were shown the data-centric explanation.

Table 5. Summary: Impact of explanation genre on consumer outcomes

Outcomes	Data-centric (control)	Features-based	Combination	Combination + rules
Performance on judgement task				
Overall	82%	-3pp***	-4pp***	-7pp***
Scenario 1: Correct Acceptance	96%	+1pp	+1pp	+0pp
Scenario 2: Correct Rejection	94%	+1pp	-0pp	-0pp
Scenario 3: Incorrect data input	92%	+2pp**	+2pp**	+0pp
Scenario 4: Overreliance on one feature	77%	-6pp***	-6pp***	-18pp***

Scenario 5: Failure to consider relevant features	51%	-11pp***	-15pp***	-20pp***
Comprehension of algorithm's decision making				
Role of algorithm	46%	-5pp**	-1pp	-12pp***
Directionality of features	86%	+2pp*	+1pp	+1pp
Features importance	67%	+18pp***	+16pp***	+18pp***
Attitudes toward the task and information provided				
…is important	93%	+2pp	+0pp	+1pp
…is helpful	92%	+2pp*	+1pp	+1pp
…is sufficient	62%	+12pp***	+13pp***	+16pp***
Confidence	72%	+6pp***	+4pp**	+4pp**

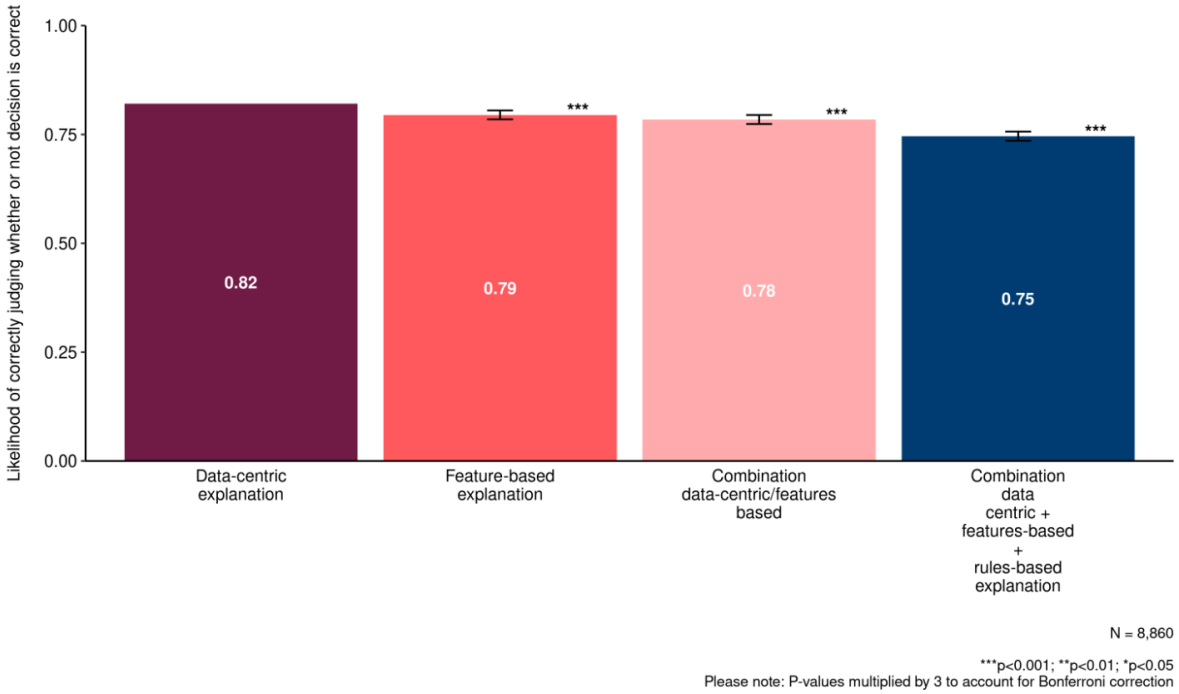
To note: * indicates significance at the 0.05 level, ** indicates significance at the 0.01 level, and *** indicates significance at the 0.001 level, including Bonferroni corrections where appropriate. We have rounded results to the nearest percentage point, which explains why graphs and tables may show minor variation in effect sizes.

Impact of genre on judgement of scenarios

Overall, providing more detailed explanations on how the algorithm works decreased participants' judgement accuracy. Participants in our control genre performed better overall at accurately judging whether a credit decision was correct or not, in comparison to those shown our treatment genres.

We first looked at participants' performance on the judgement task across all 5 scenarios tested. Participants who were shown the data-centric explanation genre had the greatest accuracy in the judgement task overall, correctly judging credit decisions around 4 out of 5 times (82%) (see Figure below). Compared to this control group, performance was between 3 and 7pp lower across our treatment genres.

Figure 3. Impact of explanation genre on judgement across all scenarios



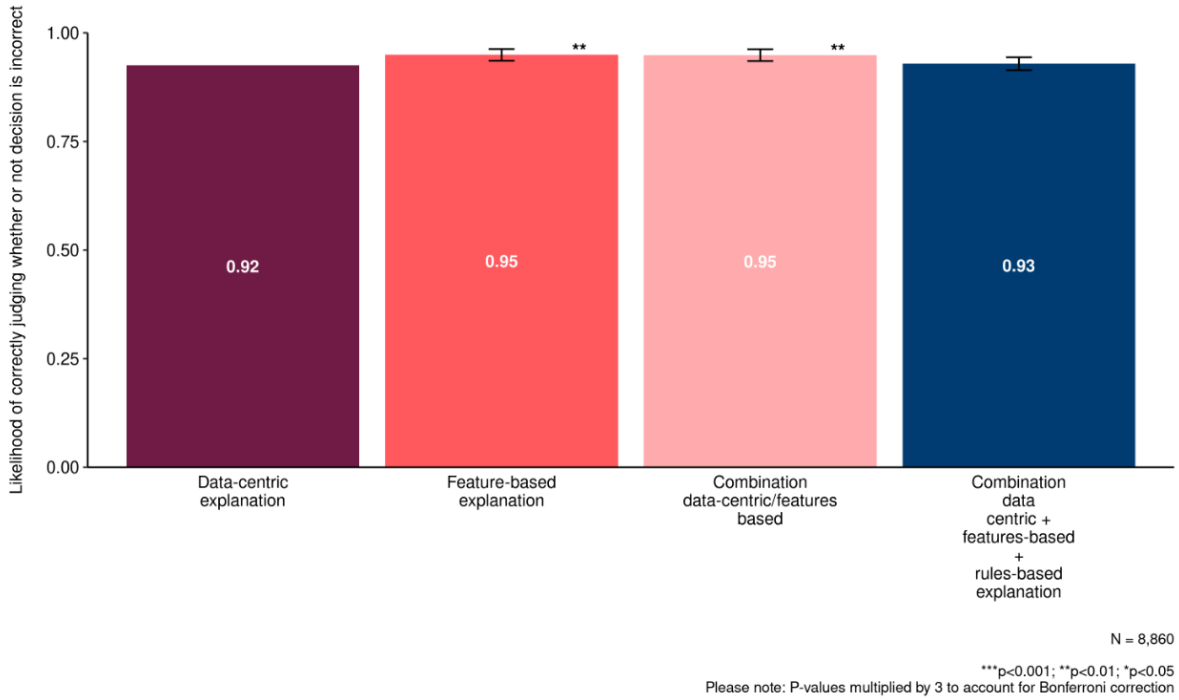
For correct credit decisions, the type of explanation given to participants did not affect their judgement - participants performed strongly across all treatment genres.

To unpick differences in average performance on the judgement task across all scenarios, we looked at how our explanation genres performed on each individual scenario. A large majority (90%+) of participants accurately judged the correct credit decisions regardless of whether the algorithm accepted or rejected the application, across all groups. We found no significant differences in performance on these scenarios between our control and treatment groups.

Participants were less accurate at detecting incorrect decisions. The explanation genres differently impacted participants' judgement of incorrect decisions depending on the type of error.

Participants detected incorrect data input (Scenario 3) approximately as frequently as they accurately accepted correct decisions, with more than 90% of participants across all groups challenging this decision. However, here we saw that participants shown the features-based and combination of features-based and data-centric explanations challenged this decision statistically significantly more than in the control, by approximately 2pp (see Figure 4).

Figure 4. Impact of explanation genre on judgement of Scenario 3 (incorrect data input)



In general, participants found it slightly harder to detect incorrect credit decisions resulting from the algorithm's overreliance on a single feature (Scenario 4) or its failure to consider relevant features (Scenario 5). In both cases we also saw worse performance among our treatment genres relative to the control. Where 77% of participants shown the data-centric explanation (our control) successfully challenged Scenario 4, participants had substantially lower performance across our treatment groups, by between 6 and 18pp (see Figure 5). Similarly, 51% of participants shown the data-centric explanation challenged Scenario 5, but performance was between 11 and 20pp lower across our treatment groups (see Figure 6).

Figure 5. Impact of explanation genre on judgement of scenario 4 (overreliance on one feature)

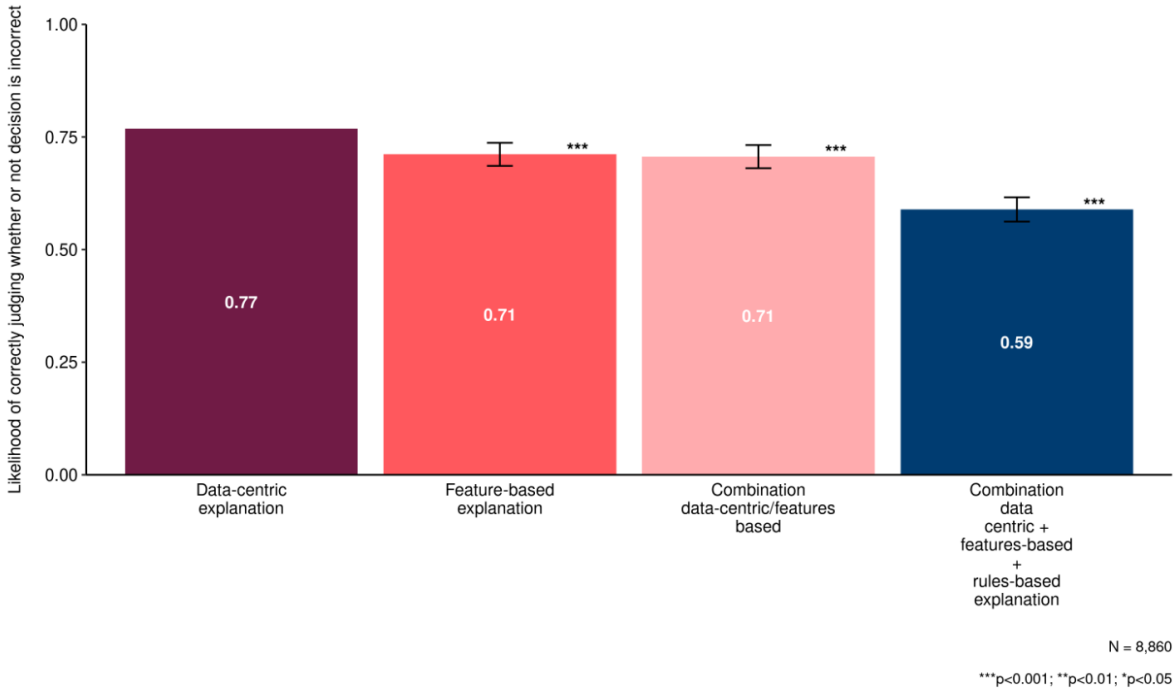
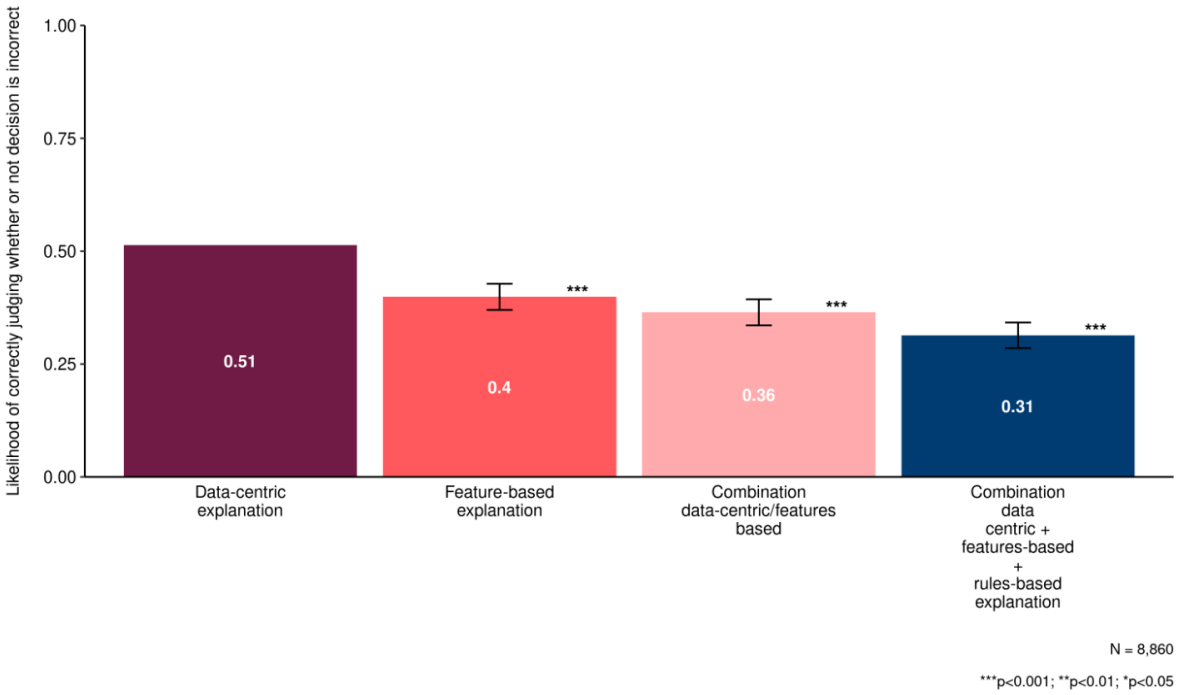


Figure 6. Impact of explanation genre on judgement of Scenario 5 (failure to consider relevant feature error)



After participants judged each of the 5 scenarios, we asked them to select a reason for their decision from a pre-populated list. While we were primarily interested in observing why participants rejected a decision, we included this question after both acceptance and rejection (with correspondingly different pre-populated lists). We report the reasons for rejection, and differences across scenarios and groups, in the Annex.

Including a decision-rule impaired participants' judgement of incorrect decisions, regardless of the error type.

Our experimental design enabled us to isolate the effect of adding a decision-rule. Adding a decision-rule to the combination of data-centric and features-based explanation significantly impaired participants' ability to accurately judge incorrect decisions, but not correct ones. For the incorrect decisions, performance ranged between 2 and 12pp lower for those shown the decision-rule than for those shown the same explanation without the decision-rule.

Table 6. Impact of adding rules-explanation on judgement of credit decisions

Scenario	Combination of data-centric and features-based explanation: % judgements correct	+ Decision rule included: % judgements correct (comparison to combination explanation)
1: Correct Acceptance	97%	97%
2: Correct Rejection	93%	93%
3: Incorrect Rejection (Incorrect data input)	95%	93% **(-2pp**)**
4: Incorrect Rejection (Overreliance on one feature)	71%	59% **(-12pp***)**
5: Incorrect Rejection (Failure to consider relevant features)	36%	31% **(-5pp***)**

To note: * indicates significance at the 0.05 level, ** indicates significance at the 0.01 level, and *** indicates significance at the 0.001 level. Additionally, the '+ Decision rule included' refers to the combination of the data-centric and features based explanation with the addition of a decision rule.

The explanation genres tested did not impact participants' ability to accurately judge credit decisions differently for different age or sex groups.

On average, individuals aged 18-24 correctly judged credit decisions 77% of the time, the lowest proportion among all age groups. The relationship between participants' age and their accuracy in judgment was consistent across all treatment conditions.

There was no notable difference between male and female participants in the proportion of credit decisions judged correctly, nor in their response to the treatments.

Impact of genre on perception and confidence

A large majority (90%+) of participants across all groups reported that the information provided to them was important and helpful, but significantly fewer participants rated the information as sufficient in the control group compared to the treatment groups.

We asked 4 attitudinal questions to assess whether explanation genre impacted participants' attitudes towards the task or information provided.

Compared to our control group (92%), we observed a marginal but statistically significant increase in the proportion of participants shown the features-based explanation reporting that the information provided was helpful (+2pp). However, there were no significant differences between the control and our other treatment groups.

Differences in the perceived sufficiency of the information provided were much more pronounced. While 62% of participants in the control group reported that the amount of information provided was sufficient, this rose by between 12 and 12pp among those in the treatment groups.

Similarly, participants in the control group were the least confident in their ability to challenge incorrect decisions (72%). Statistically significant increases in confidence were observed across our treatment groups, with 4-6pp more participants reporting confidence in their ability to challenge decisions compared to those shown the data-centric explanation.

Impact of genre on comprehension

The genre that best supported objective understanding depended on the specific question asked.

We asked 3 comprehension questions to explore whether our explanation genres impacted participants understanding of the algorithm and its decision-making. Specifically, we looked at participants' understanding of the algorithm's role, the importance of specific features in the algorithm's decision-making, and how these features influenced the likelihood of the algorithm accepting the credit decision.

A significantly greater proportion of participants shown the data-centric explanation identified the correct description for how the algorithm made the credit decision compared to those shown our treatment genres. However, notably, all groups

demonstrated poor comprehension on this question. Compared to the control group, who answered this question correctly 46% of the time, performance on this question was between 5 and 12pp worse for those shown the features-based or the combination with rules-based explanation.

Participants shown the treatment genres performed significantly better than those in the control at identifying the most important features for credit application approval. The improvement in performance on this comprehension question ranged from 16 to 18pp across the treatment groups. We expected this difference as the data-centric explanation did not include any information about the relative importance of features.

Participants demonstrated a strong understanding of the direction in which different features may impact the likelihood of a credit decision being approved. The control group established a baseline performance of 86% on this question. We found no significant differences among participants shown the combination explanation or the combination with the addition of the rules-based explanation. However, those shown the features-based explanation exhibited a statistically significant increase (+2pp) in performance.

4 Discussion

The results of our study indicate that the method used to explain algorithm-assisted decisions may significantly affect consumers' ability to challenge errors. However, our findings suggest that the most effective explanation genre may depend on the type of error. For example, some explanation genres helped participants detect incorrect data input, but in the case of other errors, the same genres misled them into accepting incorrect decisions.

The following discussion proposes 2 hypotheses for the inconsistent effects of explanation genres in our study: the salience of errors and the role of confirmatory information. Our discussion also considers the reasoning behind these hypotheses, how they connect to our other findings, the potential broader implications of our results, and questions to advance the conversation on AI explainability and its role in financial services.

The salience of errors and the role of confirmatory information

Simple errors may be easier to spot when there is less information to review
Our treatment genres were more effective than the data-centric explanation in helping participants challenge incorrect decisions caused by errors in data input. While any hypothesis is speculative, the volume of information (measured by the number of data points) provided by our treatment genres was less than that provided by the data-centric explanation. Simply, it may have been easier to detect incorrect data entry in these treatments because there were fewer pieces of information to review. This idea aligns with previous research by Poursabzi-Sangdeh et al. (2021) which found that presenting more information made it harder for participants to detect mistakes in a model.

However, this does not necessarily mean that our treatment genres are inherently better at helping consumers identify incorrect data entry, even if our hypothesis about the salience of errors is correct. Our treatment genres only describe the features considered by the algorithm in contrast to all possible features, as in the control. However, they have more columns than the data-centric explanation. Therefore, when the algorithm considers more features, the treatment genres may actually include more information. If the positive effect we observed was simply due to the ease of spotting errors, it may reverse in these cases.

It is worth noting that this is just one possible reason for why we observed differences between explanations in participants' ability to identify this error. Further research would be required to test this hypothesis.

Partially opening the 'black box' may discourage the use of personal judgement needed to challenge complex errors
Despite helping participants identify incorrect data input, our treatment genres worsened participants' abilities to challenge the algorithm when errors arose from its overreliance on, or failure to consider, certain features. We hypothesise that the treatment genres

were less effective in these scenarios because of the confirmatory nature of the information provided by them. Specifically, by highlighting the algorithm's decision logic, our treatment genres may have led participants to focus on whether the logic was followed rather than questioning whether it was sound.

The errors in these cases were not caused by the incorrect implementation of the algorithm but by how the algorithm handled certain features. Unlike data input errors, participants could not directly observe the algorithm's overreliance one feature, nor its failure to consider relevant features. Instead, participants had to rely on their personal judgement of an individual's creditworthiness given the information provided.

As these errors were more complex and required participants to evaluate conflicting information about the applicant's profile, participants may have been more likely to rely on the information about how the algorithm made its decision rather than using their own judgement to evaluate whether the decision was correct.

This hypothesis aligns with the higher rates of confidence to challenge incorrect decisions observed among our treatment groups. A more prescriptive explanation of the algorithm's decision logic allowed participants to 'figure out' the decision by relying on the decision logic, increasing their confidence in their ability to assess the decision.

The role of confirmatory information was also supported by testing the addition of the decision-rule. We found that introducing the decision-rule deteriorated participants' ability to challenge errors, substantially in some cases. This is consistent with our confirmatory information hypothesis as, by definition, the decision-rule makes the decision logic explicit. Participants may have delegated thinking about what the rule should be to the algorithm, and just viewed themselves as judges of whether it had been enforced.

What might our results indicate for the transparency and explainability of AI in financial services?

Additional information is not always helpful

The negative impact disclosure of the decision-rule had on participants' ability to challenge errors in our study demonstrates that additional information may not always be helpful. This also chimes with our hypotheses about the salience of errors and the role of confirmatory information. The salience of errors hypothesis suggests that additional information may make it more difficult to spot errors because there is simply more information to review. Likewise, the confirmatory information hypothesis suggests that additional information may encourage participants to focus on whether the algorithm's decision logic was followed rather than whether the decision logic was sound. Transparency in how AI and ML systems operate can be important. However, our findings suggest that simply providing more information may not always be the most effective way of unlocking the potential benefits of greater transparency.

Testing explanation genres in context is important

Our study shows the importance of testing explanation genres in context in order to determine the most effective information and explanations to present. As we have discussed, the impact of any given explanation genre on the outcomes measured – be it

accuracy in judgement, comprehension, or attitudes – varied depending on the specific scenario and error in question. This highlights the value of testing in context, and across contexts.

Measuring actual decision-making is valuable
This research demonstrates the value of measuring actual decision-making as well as self-reported attitudes and comprehension. Participants shown the treatment genres were more likely to report that the information provided was sufficient and that they felt confident to challenge incorrect algorithm-assisted decisions, compared to those shown the data-centric explanation. However, this conflicts with their actual performance, as they were less likely to challenge incorrect decisions on average. While measuring attitudes and comprehension may help unpick why some explanation genres are more effective than others and provide insight into participants' experiences, measuring actual decision-making might better identify how effective different approaches are.

Questions to advance the conversation on AI explainability

How can we best explain AI assisted decisions in other contexts within financial services?
This experiment only tested explanation genres in a specific context, creditworthiness, and used only a limited number of scenarios and errors in combination, for one type of financial product. Therefore, it is possible that some of the findings we observe are artefacts of the context and examples we used. For example, here we used a credit scoring algorithm, although for more complex products or decision-making, our explanation genres may have performed differently. Given the novelty of this experiment and the limited evidence in this space, we welcome further research to understand how AI-assisted decisions in other contexts, and for other financial products, can be best explained to consumers.

What mechanisms determine how changes in explanation genre impact decisions?
While our proposed hypotheses about the findings we observed are grounded in principles from behavioural science, our experiment was not designed to understand specific mechanisms for how changes in information impact consumers. Future research deliberately designed to understand how consumers process and respond to AI explainability methods could further our understanding of how best to use them to support consumer decision-making.

Can we look to other ways of presenting explanation genres?
To avoid inadvertently testing participants ability to understand different types of visual presentation, we presented all explanations using tables and standardised style across treatments. However, we're aware that some explanation genres are often presented in alternative formats in the real world. For example, the features-based explanation is often presented graphically (ie SHAP models). Alternative presentation styles may offer opportunities to enhance explainability.

Beyond explainability – how can we improve the broader consumer journey?
When thinking about how this research contributes to improved outcomes in the consumer credit space, we would encourage consideration of the broader consumer journey beyond recognising errors. For example, helping consumers understand eligibility requirements.

References

Abdi, H. (2007). Bonferroni and Šidák corrections for multiple comparisons. *Encyclopedia of measurement and statistics*, *3*(01), 2007.

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. The MIT Press.

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). 'It's reducing a human being to a percentage': Perceptions of justice in algorithmic decisions. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3173574.3173951

Dodge, J., Liao, Q. V., Zhang, Y., Bellamy, R. K. E., & Dugan, C. (2019). Explaining models: An empirical study of how explanations impact fairness judgment. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 275–285. https://doi.org/10.1145/3301275.3302310

Financial Conduct Authority. (2024, April). *Artificial intelligence (AI) update – Further to the government's response to the AI White Paper.* https://www.fca.org.uk/publications/corporate-documents/artificial-intelligence-ai-update-further-governments-response-ai-white-paper

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2017). Predictably unequal? The effects of machine learning on credit markets. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3072038

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5), 1–42. https://doi.org/10.1145/3236009

Gunning, D., Vorm, E., Wang, J. Y., & Turek, M. (2021). DAPRA's explainable AI (XAI) program: A retrospective. *Applied AI Letters*, *2*(4), e61. https://doi.org/10.1002/ail2.61

Lipton, Z. C. (2018). The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*(3), 31–57. https://doi.org/10.1145/3236386.3241340

Lundberg, S., & Lee, S.-I. (2017). *A unified approach to interpreting model predictions*. *arXiv*. https://doi.org/10.48550/ARXIV.1705.07874

Morley, J., Machado, C., Burr, C., Cowls, J., Taddeo, M., & Floridi, L. (2019). The debate on the ethics of ai in health care: A reconstruction and critical review. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3486518

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., & Wallach, H. (2021). Manipulating and measuring model interpretability. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–52. https://doi.org/10.1145/3411764.3445315

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should i trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

12 Endeavour Square, London E20 1JN
Telephone: +44 (0)20 7066 1000 Website: www.fca.org.uk