

**MASTER**

**Modifying Visual Explanations to Improve the User Understand-ability of Explainable Artificial Intelligence Systems**

Mohan, R.

*Award date:*  
2021

[Link to publication](#)

**Disclaimer**

This document contains a student thesis (bachelor's or master's), as authored by a student at Eindhoven University of Technology. Student theses are made available in the TU/e repository upon obtaining the required degree. The grade received is not published on the document as presented in the repository. The required complexity or quality of research of student theses may vary by program, and the required minimum study period may vary in duration.

**General rights**

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain

Eindhoven, May 2021



# Modifying Visual Explanations to Improve the User Understandability of Explainable Artificial Intelligence Systems

by Raghav Mohan

identity number 0804926

in partial fulfilment of the requirements for the degree of

**Master of Science  
in Human Technology Interaction**

Supervisors:

Dr. ir. M.C. Willemsen, Eindhoven University of Technology

S. Hadash MSc., Eindhoven University of Technology

Prof. dr. C. Snijders, Eindhoven University of Technology

# Abstract

Explainable Artificial Intelligence (XAI), is a domain that aims to make AI systems with increasingly complex machine learning models more transparent. In XAI, there are visualization frameworks such as Local-Interpretable Model-agnostic Explanations (LIME) that provide users with feature level explanations for individual predictions (Ribeiro, Singh, & Guestrin, 2016). We noticed that LIME explanations that are framed with respect to a negative decision class are not very easy and intuitive to understand. Although framing techniques are not systematically researched in the XAI domain, there is plenty of evidence for this in other domains such as psycholinguistics, neuroscience, and cognitive psychology (Kuhlmann, Hofmann, Briesemeister, & Jacobs, 2016; Matlin, 2016; Sherman, 1976). In this study, we investigated how two different framing techniques influenced the understandability of XAI explanations. Firstly, we manipulated the framing of feature visualizations such as those found in LIME visualizations. Secondly, we manipulated the framing of each individual feature in an explanation by including an underlying continuous variable (UCV). In many XAI tools, classifications are continuously formulated based on prediction probabilities. We thought that this could be used in the form of UCVs to improve explanations. In the current study, UCVs are framed with respect to positive and negative decision class outputs. For example, in a music AI system where a feature such as acousticness, contributes positively towards liking a particular song, the UCV for acousticness would be *+like* or *-dislike*. Furthermore, we explored whether user belief (whether the user agrees/disagrees with the prediction) and system domain influenced the understandability of our manipulated explanations (Doshi-Velez & Kim, 2017; Kahneman & Tversky, 1972). Finally, we explored if our results generalized across system domains, by testing our framing manipulations on a music AI system and a loan decision support system.

To test our predictions, we conducted a mixed-design experiment. Between subjects, we manipulated the presentation of positive, negative and no UCVs. Within subjects, we manipulated the presentation of positively and negatively framed visualizations, for both AI systems (N = 133). For each framing condition, we measured understandability, duration (the time it took to understand the explanation), and user belief over six trials. At the end each condition, we also measured perceived understandability. Each condition was tested for our loan and music AI systems.

We found that explanations containing positively framed UCVs were more understandable than negatively framed UCVs. We also found that explanations containing positively framed visualizations were the most understandable and perceived to be the most understandable. When both framing techniques were combined, we also found that explanations containing positively framed visualizations with positively framed UCVs were most understandable and perceived to be most understandable. Conversely, negatively framed visualizations with negative UCVs yielded the least understandable explanations. We found partial evidence that for positive predictions, positively framed visualizations took less time to understand than negatively framed visualizations. However, our understandability measures found that positively framed visualization were the easiest to understand for both decision classes. Based on these results, we recommend using positively framed explanations in XAI systems. We think that they are comparatively better than what is currently done in local XAI tools and might motivate more users to adapt such tools. In these tools, explanations are currently framed with respect to the positive decision class for positive feature contributions and the negative decision class for negative feature contributions.

In our exploratory analyses, we found that when users did not agree with the systems' predictions, the understandability of our UCV and visualization manipulations got influenced. We think that this co-variation with user belief is indicative of a confirmation bias. Based on our findings, we think that it even more important to use positively framed UCVs instead of negatively framed UCVs when there might be confirmation bias. Additionally, we found that positively framed visualizations worked better for our loan decision support system than our music AI system. However, application domain differences in understandability were not evident when participants first interacted with our systems. Our findings suggest that if system designers want to have similar understandability effects across domains, they should use explanations sparingly, in a contextually appropriate manner. For example, explanations can be presented in situations where users might be confused or disagree with a particular prediction.

In summary, we think that positively framed features explanations are comparatively better than methods used in current local XAI tools.

*Keywords:* XAI, explanations, visualizations, framing effect

# Contents

<b>1</b>	<b>Introduction</b>	<b>4</b>
<b>2</b>	<b>Literature review</b>	<b>5</b>
2.1	Comprehending negations in language . . . . .	5
2.2	Positivity in cognition . . . . .	6
2.3	Positive and negative explanations in XAI . . . . .	7
2.3.1	Framing effects in LIME . . . . .	7
2.3.2	Other XAI research . . . . .	9
2.3.3	Current Study . . . . .	10
2.4	The effect of application domain and user belief on understandability . . . . .	10
2.4.1	User belief . . . . .	10
2.4.2	Application domain . . . . .	11
2.4.3	Current study . . . . .	11
<b>3</b>	<b>Method</b>	<b>12</b>
3.1	Task . . . . .	12
3.2	Manipulations . . . . .	12
3.3	Trials . . . . .	14
3.4	Measuring understandability . . . . .	14
3.4.1	Understandability . . . . .	14
3.4.2	Duration . . . . .	14
3.4.3	Perceived understandability . . . . .	14
3.4.4	Valence determination task . . . . .	14
3.5	Measured independent variables . . . . .	15
3.6	Other measurements . . . . .	15
<b>4</b>	<b>Practical Implementation</b>	<b>16</b>
4.1	The music AI system . . . . .	16
4.1.1	Training data . . . . .	16
4.1.2	Machine Learning Model . . . . .	16
4.1.3	Explanation generation and visualization . . . . .	17
4.2	The loan decision support system . . . . .	17
4.2.1	Training data . . . . .	17
4.2.2	Machine Learning Model . . . . .	17
4.2.3	Explanation generation and visualization . . . . .	18
4.3	User interface design . . . . .	18
<b>5</b>	<b>Participants &amp; Procedure</b>	<b>20</b>
5.1	Participants . . . . .	20
5.1.1	Participant characteristics . . . . .	20
5.1.2	Sampling procedures . . . . .	20
5.1.3	Sample size determination . . . . .	20
5.2	Pilot Study . . . . .	20
5.3	Main Procedure . . . . .	21
5.4	Data diagnostics . . . . .	22
5.5	Analytic strategy . . . . .	23
<b>6</b>	<b>Results</b>	<b>24</b>
6.1	Descriptive statistics . . . . .	24
6.2	Principle component analysis . . . . .	24
6.3	Correlations between dependent measures . . . . .	25
6.4	Dealing with clustered data . . . . .	25
6.5	Assumption testing . . . . .	25
6.5.1	Outliers . . . . .	25
6.5.2	Normality of residuals . . . . .	25
6.5.3	Homoskedasticity . . . . .	25
6.6	Regression models . . . . .	25

6.7	Hypotheses testing . . . . .	28
6.7.1	Underlying continuous variable effects . . . . .	28
6.7.2	Visual framing effects . . . . .	29
6.7.3	Alignment between framing and prediction . . . . .	30
6.8	Interactions effects . . . . .	31
6.8.1	UCV and visualization framing . . . . .	31
6.8.2	UCV, framing and, user belief . . . . .	32
6.8.3	UCV, framing and, domain . . . . .	33
<b>7</b>	<b>Discussion</b>	<b>35</b>
7.1	The effect of framed UCVs and visualizations on user understandability . . . . .	35
7.2	Covariations in user belief, application domain and learning effects . . . . .	36
7.3	Limitations . . . . .	36
7.4	Implications and future research . . . . .	37
<b>8</b>	<b>Conclusion</b>	<b>39</b>
<b>9</b>	<b>References</b>	<b>40</b>
<b>10</b>	<b>Acknowledgements</b>	<b>42</b>
<b>A</b>	<b>Participant Instruction Screens</b>	<b>43</b>
<b>B</b>	<b>Plots for Testing Regression Assumptions</b>	<b>46</b>

# 1 Introduction

In the field of Artificial Intelligence, we see that machine learning models are becoming increasingly complex and less transparent. Industry based models that aggregate large amounts of data often use black box models that might be trading interpretability for accuracy (Tintarev & Masthoff, 2015). The field of explainable artificial intelligence (XAI) aims to increase the interpretability of all machine learning models. In fact, XAI has seen a resurgence in research publications due to this increased need for model interpretability, (Ferreira & Monteiro, 2020). In their literature review, Ferreira and Monteiro (2020) divided publications into two major research communities: The human computer interaction community, who are more focused on providing explanations to end users, and the computer science community, who are more focused on aiding technical experts. The current research deals with explanations in the human computer interaction community where users might not be so familiar with machine learning models and therefore, need explanations that are easy to digest.

For consumer based systems, explanations could be relevant in many contexts. As Miller (2019) defines it, explanations have the ability to provide an answer to why-questions. For example, in a music AI system, when a predicted song, artist or album does not entirely align with the users' expectations, they might want more information about the prediction (ie. *"why did I get this song?"*). In a loan decision support system that predicts that a user is eligible for a loan, they might want to know how the system made its decision (i.e., *"why am I not eligible for a loan?"*). One approach to make the system more transparent is to explain the entire procedure that the model uses to come up with a decision. However, explaining the entire process is not exactly what a common user with such questions wants to know (Miller, 2019). Instead they would like to understand their particular prediction. To illustrate this, consider a classifier that outputs positive and negative song predictions for a user based on their previous likes and dislikes. In a music AI system this could be a prediction such as *"The system thinks that you dislike this song"*. Local Interpretable Model-agnostic Explanations or LIME, has the ability to provide explanations for such predictions. In particular, LIME provides weighted featured explanations that contribute to each decision class labels to explain local predictions (Ribeiro et al., 2016). So for a music AI system, LIME could show how song features such as acousticness and danceability contribute towards the *like* and *dislike* decision classes, to explain why the system predicted *like* or *dislike* for a particular song.

The problem occurs when the model contains classes which users inherently interpret as being positive/negative. In particular, if a prediction is negative and the explanations are framed with respect to the negative decision class, the feature explanations become difficult to interpret. For example, the explanation *"Credit history contributes negatively towards not getting a loan"* takes longer to process than *"Credit history contributes positively towards getting a loan"*. This example demonstrates that negatively framed explanations could hinder a user's ability to understand an output. In other domains such as psycholinguistics, it has been shown that people generally have difficulty understanding negatively framed sentences (Sherman, 1976). In fact, in cognitive psychology it has been shown that people initially accept all incoming information before deciding what to reject (Gilbert, Krull, & Malone, 1990) and our cognitive processes handle positive information better than negative information (Matlin, 2016). Even though the XAI domain emphasizes that explanations should be written in a unambiguous manner (Miller, 2019), and although these difficulties in understanding have been observed (Stumpf et al., 2007), there has not been a systematic evaluation of the effect of negatively and positively framed explanations on user understandability.

Therefore in the current study, we investigated how positively and negatively framed explanations contribute towards providing end-users with a better understanding of local feature level explanations. To show whether our investigations can be applied to multiple domains, we explored framing effects in our own music AI system and loan decision support system. Additionally, people might not always agree with a prediction and this might influence their ability to understand feature explanations. Therefore, we also explored if misaligned user belief and system predictions can influence the ability to understand such explanations. Finally, we examined how presenting similarly framed explanations multiple times could influence explanation understandability. If users can understand positively framed explanations better, these explanations can potentially help system designers convert explanations generated by local XAI tools such as LIME, into explanations suitable for end-users.

## 2 Literature review

### 2.1 Comprehending negations in language

In psycholinguistics literature, it has been found that people take longer to process sentences containing words with negations such as “no”, “not”, or “neither”. Sherman (1976) found that comprehension decreased when sentences with two or more negations required processing. So for example “no one” has one negation while “no one doubted not” has three negation. Sherman (1976) found that both error rates and response times increased with more negatives. Negations can also be reformulate into positive equivalents. In fact, Corblin (1996), suggests that people take a longer time to understand a sentence with two or more negations compared to a semantically equivalent sentence with positive words. For example, the sentence “I don’t deny that I am not guilty” can be translated into the semantically equivalent sentence “I am guilty”. Another study by Kaup, Lüdtke, and Zwaan (2006) investigated if people have mental representations of sentences containing such double negations. In their study, participants first read a sentence (e.g. “The door was not open”) and were subsequently presented with a picture that either matched (closed door) or mismatched (open door) the sentence. Kaup et al. (2006) found that when the picture was presented shortly after a negated sentence (750ms), participants chose the mismatched picture. However, when they had a longer time between a negated sentence and a picture (1500ms), the effect was mitigated. For our study, the study by Kaup et al. (2006) implies that mental depictions of negatively framed explanations might only be incorrect for short periods of time after which they might learn the negations. Lenzner, Kaczmarek, and Lenzner (2010) also found this effect in survey designs, where double negated survey items had a longer processing time (e.g. “Was this facility not unclean?” vs “Was this facility clean?”). Overall, all these studies provide us with initial evidence that explanations with multiple negations might be hard to comprehend.

Other than the valence of individual words, truth values of sentences might also influence processing time (Kess, 1992). For example, “Toronto isn’t above Texas” is easier to verify than “Texas isn’t above Toronto”. This is because in the former sentence, the negation “isn’t”, aligns with the truth value of the sentence (false). For the second sentence however, the negation “isn’t” does not align with the truth value (true). This was found for implied truths (e.g. “It’s true that the red dots are red” vs “It isn’t true that the red dots aren’t red”) and truths with visualizations (e.g. Star is above plus  $^+*$  vs Plus isn’t above star  $^+*$ ) (Carpenter & Just, 1975; Clark & Chase, 1972). In neuroscience literature, studies found that brain areas related to semantic violation, semantic reintegration, stimulus processing, language and, visuospatial processing were activated when truth values were misaligned with word valences (Carpenter, Just, Keller, Eddy, & Thulborn, 1999; Fischler, Bloom, Childers, Roucos, & Perry, 1983; Herbert & Kübler, 2011; Kutas & Federmeier, 2000). The neuroscience studies show that when semantic violations are combined with negations, people pay more attention to the sentence. There is fundamental evidence for this in our brain. From these studies we see that when a truth values is misaligned with a sentence valence it decreases comprehension in many situations. In our study we test if this effect holds true for positive and negative predictions. In particular, we hypothesize that if the valence of a prediction does not align with the valences in our feature explanation, people might have a hard time comprehending the explanation.

Many of these studies looked at the misalignment between word valences and sentence truth values. Although this is partially relevant for our study, they did not investigate misalignment between word valences and sentence valences. In our study, feature explanations are generally true, but there is incongruity between the valence of the overall explanation and individual words in the explanation. In the current study we ask questions such as: In a loan decision support system, is the feature explanation “Your saving status (<\$100) contributes positively towards not getting a loan” harder to comprehend than “Your saving status (<\$100) contributes negatively towards getting a loan”? Kuhlmann et al. (2016) investigated a similar question on the word level. They examined human processing times for German bivalent compound words where the valences were either both positive “Erotikengel” (erotic angel), both negative “Arrestschuft” (arrest villian), positive-negative “Diplomübelkeit” (diploma nausea) or, negative-positive “Konfliktoptimist” (conflict optimist). In their study, participants were placed in an fMRI scanner and asked to perform a valence determination-task on 120 such words. In the current study, we used a similar valence determination task to check if people were paying attention. In their valence determination task, participants were asked to categorize if a given bivalent word was positive or negative within a fixed time frame. Kuhlmann et al. (2016) found that participants reacted faster when valences were aligned. Another study, which provides the closest evidence for the current study, investigated misaligned valences on the phrase level with word pairs (Itkes & Mashal, 2016). In the first experiment, they investigated pairs where the valence of overall word pairs was aligned with the first

word (e.g. “good dog”). In the second experiment, they investigated misaligned combinations between first-words and word-pairs (e.g. “blood relation”). In both experiments, [Itkes and Mashal \(2016\)](#) found that response times were faster when the valence of the overall word pair was positive, even if the first word was negative. These results provides initial evidence for our own study that people might be quicker in processing positively framed explanations.

In psycholinguistics and neuroscience, we have presented a lot of evidence that people have a hard time comprehending sentences with word valences that are misaligned with the truth and/or valence value of phrases and sentences. However, to see why positively framed explanations are particularly easy to understand for human beings, we look at theories from cognitive psychology.

## 2.2 Positivity in cognition

[Itkes and Mashal \(2016\)](#) were actually comparing two cognitive theories in their study. They wanted to see if their results provided evidence for the Affective Primacy Hypothesis or the Cognitive Primacy Hypothesis. The Cognitive Primacy Hypothesis states that the mental process of understanding semantic meaning, precedes affective activation ([Lazarus, 1984](#)). The Affective Primacy Hypothesis states that affective information has priority ([Zajonc, 1980](#)). In both their experiments, [Itkes and Mashal \(2016\)](#) propose that for negative first words, valence was only activated after understanding the entire word-pair. Therefore for negative valences, they provided evidence for the Cognitive Primacy Hypothesis. This suggests that positive information might occur earlier on during the processing stage. However, the Cognitive Primacy Hypothesis might not hold for all contexts. In fact, the study by [Lai, Hagoort, and Casasanto \(2012\)](#) found that either hypotheses might hold depending on the stimuli and context. If people were placed in a highly affective context such as seeing a picture of a snake, they found that affective judgements were faster than non-affective judgement thereby providing evidence for the Affective Primacy Hypothesis.

Another study compared two similar ideas by investigating if understanding and belief co-occur or if they are separate processes ([Gilbert et al., 1990](#)). They aimed to resolve an age-old debate between philosophers Rene Descarte and Baruch Spinoza. While Descarte argued that people take time to understand and believe all information before making a final decision (reject or accept), Spinoza argued that all information is initially accepted after which people decided if they wanted to reject things. For the primary study conducted by [Gilbert et al. \(1990\)](#), participants were asked to sentence criminals based on robbery statements which varied in severity. In some of the trials (interruption trials), participants were provided with green text indicating true statements or red text indicating false statements. If Spinoza’s theory was true, then participants will mistake the interruption trials with false statements to be true since they initially take everything to be true. If Descartes’ theory was true, subjects would not be influenced by these false statements. The results found that in interruption trials, participants mistook the false statements to be true which supported Spinoza’s theory. In the other studies by [Gilbert et al. \(1990\)](#), they tested these hypotheses for tasks involving sentences, pictures and overall comprehension. Therefore, the studies by [Gilbert et al. \(1990\)](#) provide evidence for Spinoza’s theory.

In the context of our current study on XAI, we expect that the Cognitive Primacy Hypotheses and Spinoza’s theory are in play. After receiving a prediction, we think that people place more importance in understanding the semantics behind an explanation. We think that understanding the interruption trials in the study by [Gilbert et al. \(1990\)](#), can be somewhat likened to understanding positive and negative words in an explanation. Therefore, while trying to understand the semantics of an explanation, if people are presented with explanations containing negative words, they might have a larger chance to misinterpret its meaning compared to an explanations with positive words. For example, if a music AI system predicts that you like a song and a user glances through an feature explanation such as *danceability contributes negatively towards disliking the song*, they might mistake *disliking* to mean that danceability contributes negatively towards their prediction.

Other than these two theories, a running theme in cognition is highlighted by the Pollyanna Principle which states that people typically process positive items more accurately and efficiently than negative or neutral items ([Matlin, 2016](#)). Although this is now accepted to be a recurrent theme in memory and judgment, one of the earliest evidence supporting this principle was found in a language based study where they investigated how word valency influences processing time ([Boucher & Osgood, 1969](#)). They found evidence across thirteen cultural communities that positive words were used more frequently than negative words. In eleven of these communities, people used negative affixes (such as un-, non-, dis-) for positive words (unhappy) more frequently than negative affixes for negative words (unsad, a double negation). However, the Polyanna Principle might not hold for negative categories where negative items



are listed first (Silvera, Krull, & Sassler, 2002). For example, if people are tasked with naming items belonging to the word “pandemic”, they most likely start by mentioning negative items. For our study, the Polyanna Principle provides some support that positively framed explanation might be easier to understand than negatively framed explanations.

The Cognitive Primacy Hypothesis, Spinoza’s theory and the Polyanna Principle all show that people tend to handle positive information with less chance for errors and with lesser difficulty. In the current study we wanted to examine if these theories carried over to positively framed feature explanations in the XAI domain.

## 2.3 Positive and negative explanations in XAI

### 2.3.1 Framing effects in LIME

Machine learning models in the industry which use large datasets and require accurate outputs, often choose to use black box models. In a black box model, after a learner is trained on a data set, a person is provided with information such as the in-sample accuracy and a confusion matrix. The person can also use test data sets to determine out-of-sample accuracy and check for over-fitting. If a model is openly accessible, the person might also be able to input various data and receive predictions as outputs. However, the person can not use predictions to directly understand if an underlying model makes sense in real life scenarios unless the feature weights and directionality for individual data points are explained. For consumer interfaces such as music recommender systems, feature weights and directionalities could help users get an explanation to potentially answer why a particular song was recommended to them. To help solve this problem, Ribeiro et al. (2016) devised such a framework with Local Interpretable Model-agnostic Explanations (LIME). With LIME it is possible to observe a particular data-point and investigate how underlying features contribute to the output value of the data-point. To illustrate such an explanation, we inputted a loan decision support system that was trained using a random forest classifier, into LIME. Figure 1 shows the visualization generated by LIME for a single loan applicant.

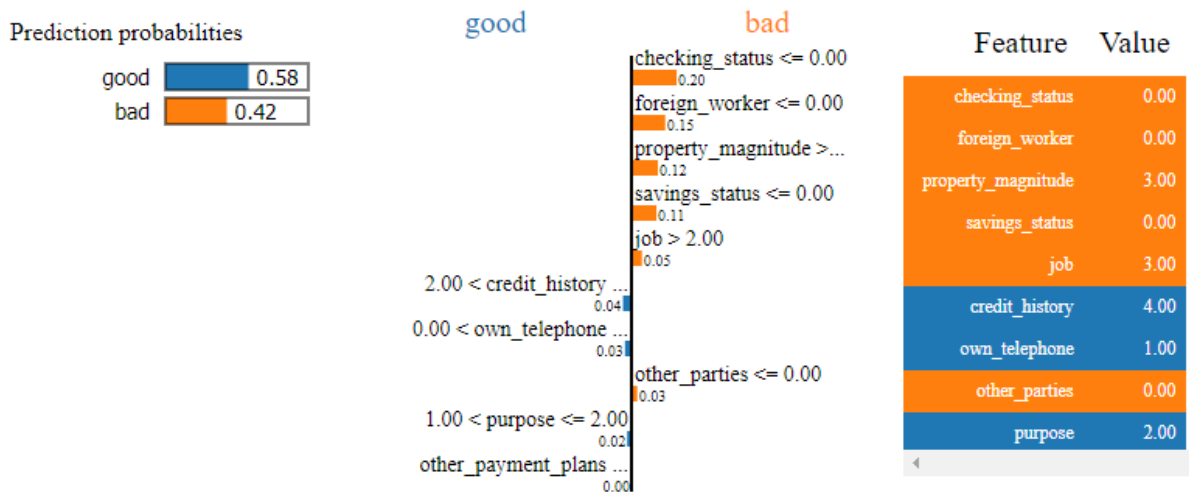


Figure 1: Lime explanation visualization with the prediction probabilities (left), bar visualization (middle) and the feature values (right)

We noticed two problems with this explanation. Firstly, this explanation uses prediction probabilities. Miller (2019) argues that prediction probabilities don’t provide useful explanations because people generally have a cognitive bias towards discounting probabilities. Secondly, it is hard to see that the final prediction is *good* from the bar visualization in figure 1. We see that each feature explanation is displayed with respect to how it contributes to a particular decision class. So for example, checking status contributes positively towards the *bad* decision class while credit history contributes positively towards the *good* decision class. When the final decision is positive, it seems odd to frame feature contributions with respect to the negative decision class. Instead, we think that the explanation should be framed with respect to the *good* decision class since people are trying to understand why the system chose *good*. After all, in our situation people want to understand the semantics behind our prediction rather than

its affective information (Itkes & Mashal, 2016). LIME explanations become even harder to comprehend when the final decision class is negative. Suppose that the decision is *bad* and credit history still contributes towards *good* while checking status still contributes towards *bad*. Now since people are trying to understand why they received the *bad* prediction, they might interpret credit history as "*Credit history contributes negatively towards bad*" which is the same as saying "*Credit history contributes positively towards good*". This is similar to psycholinguistics literature where it was found that people take longer time to process a double negation compared to its semantically equivalent positive (Corblin, 1996; Kaup et al., 2006). So even when the decision class is *bad*, it seems to be better to frame explanations with respect to the positive decision class. Therefore, in this example, regardless of decision class, it might be better to say that credit history contributes positively towards good while checking status contributes negatively towards good. The reasoning also aligns with the experiments conducted by Itkes and Mashal (2016) where they found that positive word pairs were faster to understand compared to negative word pairs, regardless of the valence of the first word. Positive framing might also makes sense from a cognitive standpoint since it was found that people tend to have less difficulty handling positive information (Matlin, 2016).

For these reasons, we explored the second problem with figure 1 more systematically by testing the effectiveness of framing approaches on understandability. First of all, in many XAI tools, we observed that classifications are continuous. For example in figure 1 we see a 58% prediction probability for *good* and a 42% prediction probability for *bad*. We thought that this continuity can be used to improve explanations. Based on this, our first approach makes use of framing an underlying continuous variable (UCV) that textually frames every single feature with respect to inherent decision class labels. In figure 1 the UCV would be positive (*loan eligible*) or negative (*loan ineligible*). In a music system these would be *like* and *dislike*. Our second approach makes use of framing visualizations where we frame the entire bar visualization with respect to a positive decision class (*loan eligible*) or a negative decision class (*loan ineligible*). We do this by changing the color and the directionality of the bar visualizations in figure 1.

To further elaborate on these framing techniques, let's look at a toy example of the LIME explanation in figure 1 where the system predicts *loan eligibility*. In table 1 we see all the framing possibilities for *property magnitude* and *credit history*. Each row represents different possibilities for our UCV while each column represents a different possibilities for framing our visualization (using colors). The LIME explanation in figure 1 is somewhat similar to the middle right cell in table 1. Look at this cell, we see that the main difference is that *framing* is worded positively with *Blue contributes towards eligibility*. Positive framing already makes this example slightly easier to understand than the LIME visualization. If such an explanation is framed with respect to the negative decision class however, it becomes quite difficult to understand since the prediction is positive. For example, in table 1 the middle left cell difficult to understand because we have to translate the wording (*blue contributes towards ineligibility*) and apply it to the feature contribution. Since the user wants to understand why the system predicted *eligible*, they could reason: *Because credit history is colored orange, it contributes negatively towards ineligibility which therefore means that it contributes positively towards eligibility*. Now, if we also add a negatively framed UCV to the negative framing such as the top right cell in table 1, we have to process two negations. In this case, a user could reason: *We see that credit history contributes negatively towards ineligibility. It is colored orange which also indicates negativity towards ineligibility. Therefore they both mean that credit history contributes positively towards eligibility*. If a user has to reason like this for multiple features, it becomes cumbersome. The amount of negations that a user has to reason through increases when we add a negative UCV which might decrease comprehensibility (Sherman, 1976). Even while going through the toy example we see that when negations are present in the UCV, it is important to pay close attention and we possibly verify our thoughts with *framing* (Gilbert et al., 1990). In contrast, we think that when we add a positive UCV to the explanation, it reinforces understandability because it not only takes less time to process positive information, but the information is directly available for each feature (Gilbert et al., 1990; Matlin, 2016).

**Table 1** Toy example with all ucv and color framing possibilities

	<b>negative framing (blue contributes towards ineligibility)</b>	<b>positive framing (blue contributes towards eligibility)</b>
<b>negative ucv</b> (+/- ineligible)	property_magnitude: +0.12 ineligible credit_history: -0.04 ineligible	property_magnitude: +0.12 ineligible credit_history: -0.04 ineligible
<b>no ucv</b>	property_magnitude: 0.12 credit_history: 0.04	property_magnitude: 0.12 credit_history: 0.04
<b>positive ucv</b> (+/- eligible)	property_magnitude: -0.12 eligible credit_history: +0.04 eligible	property_magnitude: -0.12 eligible credit_history: +0.04 eligible

*Note: In our practical implementation, we use colored bars relative to the weight instead of colored text.*

Although framing might influence positive and negative predictions, a classification model need not have a valence. Predictions can also be neutral “*this ball is red*”. The same logic can be extended to a neutral prediction if the problem is rephrased. Taking the example “*this ball is red*”, we might have a feature called color and the model might be inputted into a local XAI tool like LIME. When a *ball* datapoint contains the feature *color* that is equal to *red*, it contributes *positively* towards classifying the ball as red. However, if the datapoint contains another *color* value, it contributes *negatively* towards classifying the ball as red. Therefore, a neutral decision class can be rephrased to contain a valence (a ball being red is positive, while a ball being another color is negative) and so can its explanation (the color feature being red contributes positively towards the ball being red). In the current study however, we limit the scope to systems with inherently positive and negative decision classes.

### 2.3.2 Other XAI research

In the XAI domain, explanation quality is considered to be quite important. As Miller (2019) states, for effective human communication, explanations in XAI should adhere to Grices Maxims:

1. Quality: Information should be of high quality.
2. Quantity: Make sure that all the required information is provided and not more.
3. Relation: Only provide relevant information.
4. Manner: Avoid obscurity of expression, ambiguity and be brief and orderly.

When an explanation contains negatively framed words or UCVs, there is some level of obscurity which clearly violates the fourth maxim. In XAI literature, there is one evidence indicating that negative explanations might be hard to understand. In a qualitative study Stumpf et al. (2007) used an email classification algorithm to split an inbox into categories such as “*personal*”, “*resume*” or, “*bankrupt*”. Subsequently, participants were randomly provided with emails containing different types of explanations and they had to determine if the classifier correctly categorized the emails. In order to do this task it was important that participants understood the explanations. In one explanation condition, participants received keyword-based explanations where each email was scanned for keywords. The explanations displayed the top five keywords that contributed positively towards the categorization and the top five keywords that contributed negatively. In this condition, participants found it hard to understand how the negative keywords contributed to the overall classification.

Although not dealing with the explanation framing, Cramer et al. (2008) measured how explanations influenced the understandability of a system. In their study, participants interacted with a recommender system and chose six art pieces where system transparency was manipulated by adding or removing explanations for recommendations. In this study *actual understandability* was measured through an interview question: “*Could you please tell me how the system works? It is fine if your explanation turns out not to be accurate. This question is not to test you.*”. Cramer et al. (2008) found that transparency with the use of explanations, influenced system understandability. Cramer et al. (2008) also measured perceived understanding with a questionnaire. Interestingly, they found that it did not correlate with actual understanding which could imply that participants did not understand the system as well as they thought.

In the current study, we also measured understanding and perceived understanding. However, we measured the understanding of a given explanation rather than system understanding. In our context, users typically require explanations for local instances rather than entire systems.

### 2.3.3 Current Study

In this section we argued why LIME visualizations such as figure 1 were not easy to understand based on literature found in cognitive psychology and psycholinguistics (e.g. Gilbert et al., 1990; Sherman, 1976). With table 1 we provided examples on how these visualization could be improved with framing techniques. Although we found no XAI literature that systematically dealt with such framing techniques, we did find some evidence that framing does influence understandability (Stumpf et al., 2007). Therefore, in the current study we investigated framing techniques by answering the following research questions:

**RQ1:** How do feature explanations containing underlying continuous variables influence the ability to understand explanations?

**RQ2:** How does a framed visualization with feature explanations and its valence alignment with the final prediction, influence the ability to understand recommendations?

We proposed the following hypotheses based on the arguments made in the current section:

**H1a:** Feature explanations with underlying continuous variables will be easier to understand than explanations without underlying continuous variables.

**H1b:** Feature explanations with positive underlying continuous variables will be easier to understand than explanations with negative underlying continuous variables.

**H2:** Feature explanations with positively framed visualizations will be easier to understand than explanations with negatively framed visualizations.

Additionally, recall that in section 2.1 that literature from psycholinguistics found evidence that people might have a hard time comprehending sentences where the valence of individual words are misaligned with the sentence truth value (e.g. Kaup et al., 2006). We expected a similar effect when the valence of a prediction was misaligned with its feature explanations.

**H3:** Feature explanations where the visualization framing and prediction are aligned will be easier to understand than explanations where the visualization framing and prediction are misaligned.

Although framing might influence understandability, the understandability of a framed explanation might also be influenced by other factors.

## 2.4 The effect of application domain and user belief on understandability

### 2.4.1 User belief

All human beings have cognitive biases and heuristics that might help or hinder their ability to judge situations. One such bias is the confirmation bias where an individual only recalls information that validates their prior beliefs. Confirmation bias can lead towards potentially inaccurate judgements (Kahneman & Tversky, 1972). If a person keeps validating their own beliefs through various instances in life, this starts to become their subjective reality. For example, a user who listens to a song playing in the cafe might really enjoy the song and add it to an AI system to try and understand why the song was so enjoyable. However, this song might not align with anything that they have listened to in the past. If this system is a classifier, it might therefore predict that they don't like this particular song. Any explanation that tries to convince the user that they don't like the song, might not be effective because the user disagrees with this prediction. In fact the user might not even want an explanation in this case.

Although not systematically investigated, misaligned user belief was found to be problematic in one XAI study. In this study, researchers trained an e-meter which predicted the emotional valence of a text by classifying words in the text as positive, negative or neutral (Springer & Whittaker, 2019). Participants were required to input their own written text with an emotional valence. In one condition (the transparent condition), participants were directly provided with information on how each word in their personal text was classified (positive, neutral or negative). In this condition, one user thought that the word "isolation" would greatly contribute towards a negative valence but the e-meter actually marked the word as unimportant. Therefore in this case, the participants' belief of the word (negative), was not aligned with the model classification (neutral). In this study, user trust decreased with transparency. Participants also felt that the model was more error prone in the transparent condition. Interestingly, this study conflicts with previous research on transparency (Cramer et al., 2008). The music example and the e-meter study both illustrate that user belief might influence the need for a feature explanation. If the need for a feature explanation is less, users might pay less attention to the explanation which in turn might influence the understandability of the feature explanation.

### 2.4.2 Application domain

In their paper, [Doshi-Velez and Kim \(2017\)](#) talk about incompleteness of machine learning problems. Unlike uncertainties, incompleteness refers to a problem that creates some level of *unquantified bias*. They hypothesized that various application domains have similar or different levels of incompleteness. For example, when the goal of a system is to provide users with song predictions, there might be environmental, social and temporal variations that could potentially influence listening habits. Some of these factors might be outside the control of a music AI system. For example, when a user goes to many parties, they might listen to more upbeat music for dancing. A person who plays the violin might listen to a lot of classical music. The user's social circles and the content they are exposed to outside the system might also change their tastes. Listening habits could also be influenced by moods such as gloominess, cheerfulness, melancholy and, romanticism. Musical taste might also change over time as music proficiency grows. For a loan decision support system however, the problem is relatively more complete. There is only one decision that needs to be made (loan eligible or ineligible) and banks usually have sufficient personal and financial information to get a reasonable understanding of the situation. Since a loan decision support system is addressing a more complete problem, the explanations might be more understandable for many users. This is because the features for a loan decision support system might be less subjective than music AI systems.

The need for an explanation might also differ based on domain importance. There is a lot at stake while deciding to accept or reject a loan application since it can potentially change an applicants financial situation. For the loan applicant, an explanation could be very helpful in verifying if any incorrect conclusions were made during the decision making process. Explanations might also be important for a loan expert trying to evaluate a decision support system with data-points containing loan applicants. In this case, explanations are important since judgement made by the loan expert affect many more loan applicants. To verify if a system is providing correct predictions, a loan expert might require more fine-grained feature explanations than a loan applicant. In contrast, music predictions might be used for entertainment purposes and have a relatively lower impact on daily life. Based on this reasoning, a loan decision support system might have a greater need for explanations than a music AI system.

Therefore, across application domains, explanation understandability might change based on the completeness of a problem and the importance of a particular domain.

### 2.4.3 Current study

Based on the observation made in this section, we conducted further exploratory analyses on how user belief and application domain influence the understandability of our framed explanations. For user belief, we wanted to see if our findings changed when users disagreed with our system predictions. For application domain, to see if our findings can be generalized across two different application domains, we created a music AI system and a loan decision support system and tested our framing effects.

### 3 Method

The upcoming sections describe our experimental manipulations, dependent measures, and independent measures.

#### 3.1 Task

In our study, each participant went through multiple trials where they interacted with our system. As shown in figure 2, for each trial participants received a prediction that was positive or negative along with the main explanation which consisted of a contribution (feature weights), property (feature) and value column (feature value). The color and magnitude of each bar indicated how a particular feature value contributed towards the final prediction. To investigate how framed explanations influenced the understandability of our explanations, we manipulated framing in the contribution column and asked participants how well they understood the explanation in this column.

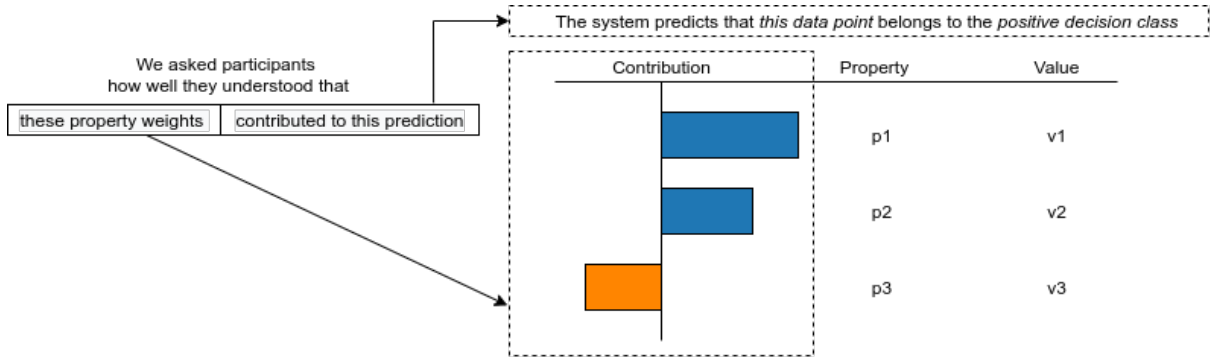


Figure 2: The primary task for each trial. Note: The bar visualizations were inspired by LIME visualizations like figure 1.

#### 3.2 Manipulations

We used a 3x2x2 mixed design where we manipulated the contribution column in figure 2 according to our hypotheses. Our framing manipulations for figure 2 can be seen in figure 3. The figure shows all six possible framing combinations for a positive prediction. These same combinations also apply for negative prediction.

To test our first hypothesis, we manipulated the underlying continuous variable or *UCV*. Each row in figure 3 is a separate *UCV* condition. Out of the two framing manipulations, we expected that *UCV* will have the largest effect size because we are manipulating framing for each individual feature. Comparing across *UCV* conditions in figure 3, we can see that is difficult to manipulate *UCV* within participants without directly revealing the conditions. Therefore, we manipulated *UCV* between participants. Participants were either presented with *UCVs* framed with respect to the positive decision class (*pucv*), negative decision class (*nucv*) or no decision class (*nucv*). Table 2 shows how participants were distributed between *UCV* conditions.

**Table 2** Distribution of *UCV* conditions.

<i>UCV</i>	n
nucv	42
nucv	42
pucv	49

*Note: We see more participants in the pucv condition because we used simple randomization since equal distribution is not a requirement for the multi-level regression we performed.*

To test our second hypothesis, we manipulated *visualization framing*, where each column in figure 3 represents a condition. Although we expected that visualization framing would also have a substantial effect based on the finding from previous literature (e.g. Itkes & Mashal, 2016; Kuhlmann et al., 2016),



manipulating two variables between participants would greatly reduce our power. So for practical purposes, we manipulated *visualization framing* within participants. We manipulated our visualizations by changing the wording above each explanation (eg. *blue contributes towards the positive decision class*), the color and, the directionality of each bar. To prevent order effects, we counterbalanced the two visualization framing conditions.

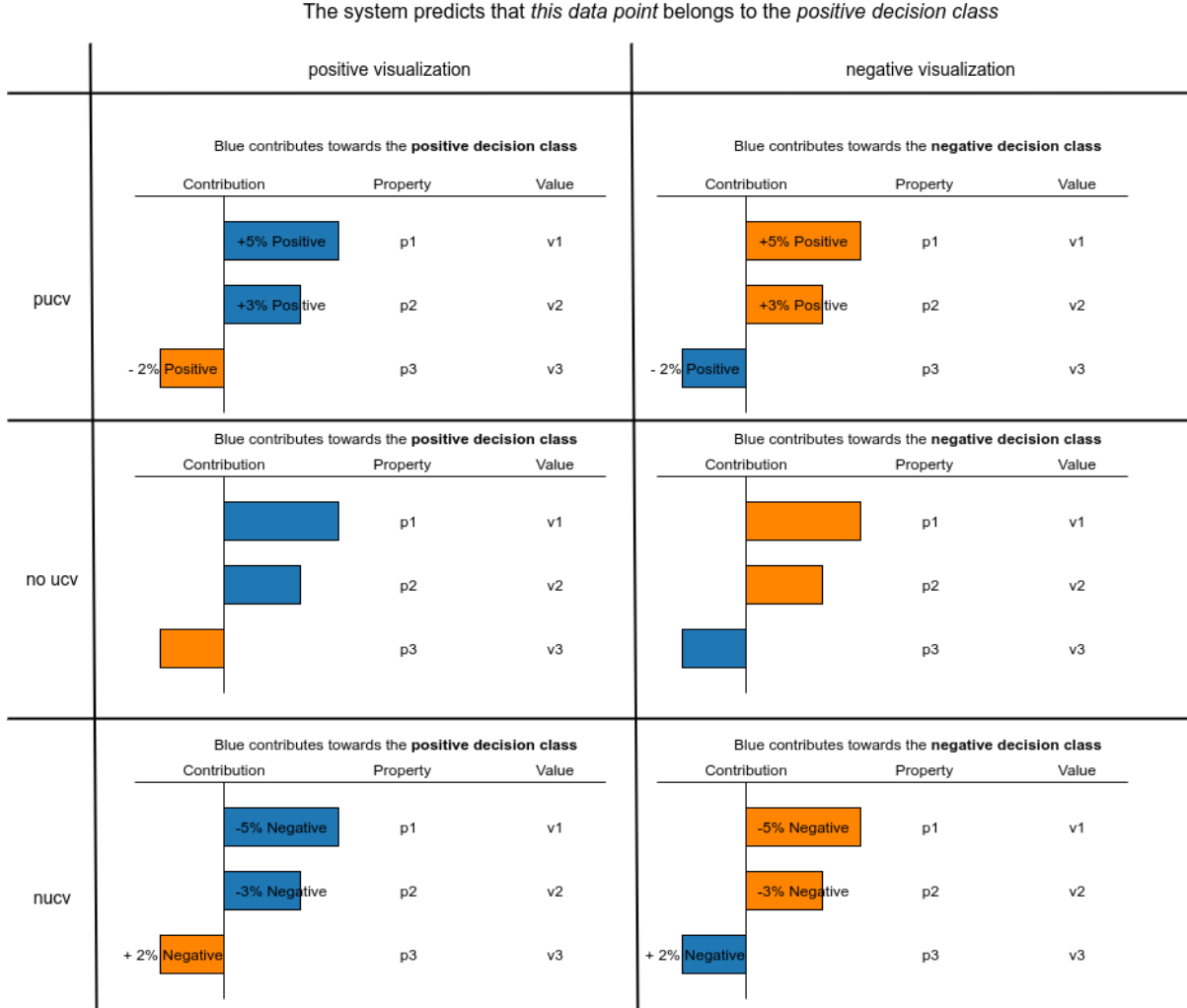


Figure 3: Our framing manipulations. Note: We manipulated the UCV conditions (rows) between participants and the visualization (columns) within participants. Also, since we are already manipulating framing with wording, we use orange and blue bars to prevent an additional manipulation.

To test our third hypothesis we also controlled how many positive and negative predictions were in each visualization manipulation. We manipulated this within participants because we did not expect a substantial effect for our third hypothesis.

To explore if our explanations generalized across domains, participants interacted with our loan decision support system and music AI system. The user interfaces for the music and loan system in figures 8 and 9 respectively and will be discussed in more detail in section 4.3. Although there are some differences, both systems follow the same main-explanation formats as figure 3. To prevent order effects, we counter-balanced *domain* across framing conditions. However, in order to avoid confusion across the different framing conditions, explanations in the music AI system were separated from the loan decision support system system.

### 3.3 Trials

We tested each framing condition across six trials to increase the precision of our measurements and to investigate any potential learning effects that might occur after their first impressions. To control for decision class, three out of six trials contained positive predictions while three contained negative predictions. Each participant went through one *UCV* condition over two *visualization framing* conditions which is represented by each row in figure 3. They repeated both conditions for both application domains. Therefore, each participant completed 24 trials. For a more detailed participant flow chart, see figure 10.

### 3.4 Measuring understandability

Explanation understandability was estimated using three different measures. In the upcoming sections we describe each measure. We then describe our valence determination task which was inspired by Kuhlmann et al. (2016) and checks if participants were paying attention in each condition.

#### 3.4.1 Understandability

Our main dependent variable was *understandability*. We define *understandability* as how well participant think they understand an explanation. In our study, after each trial, *understandability* was measured as the response to the following question:

*How well do you understand the explanation of how the properties contribute to the prediction (i.e. the contribution column)?*

*Scale: Not at all - Somewhat - Mostly - Completely*

To see how this question looks in our user interface, see figure 8.

#### 3.4.2 Duration

To complement our subjective measurement for understandability, we measured *duration*, or the time it took for people to answer the understandability question (in millisecond). We measured *duration* for each trial. In our previously mentioned psycho-linguistic studies, the most commonly used measure for understandability was the time that it took for participants to understand a sentence (Clark & Chase, 1972; Kaup et al., 2006; Lenzner et al., 2010; Sherman, 1976). We expected that feature explanations will take a longer time to understand than sentences and therefore, might be subjected to larger variations. To lessen this effect, participants were asked to provide a response as soon as they had an answer in the instruction screen prior to the trial. The participant instructions for each application domain can be found in appendix A.

#### 3.4.3 Perceived understandability

After six trials participants might have a better overall perception of a given condition. Therefore it makes sense to have an measurement for *perceived understandability*. *Perceived understandability* measures how participants feel about each condition. For each participant, this was measured four times (twice for each domain) using the following items:

*To what extent do you agree with the following statements:*

- *The explanations helped me get more insight into the given prediction.*
- *The explanation felt clear to me.*
- *I felt that the explanations took me a lot of time to comprehend.*
- *The explanations were confusing to me.*

*Scale: Strongly Disagree - Disagree - Neutral - Agree - Strongly Agree*

#### 3.4.4 Valence determination task

The valence determination task or *vd*t, provided an additional check to see if participants understood the explanations. After each framing condition, one of the previous trials was chosen at random. Figure 4 shows a *vd*t for a trial from our music AI system. As shown in the figure, a prediction and explanation were displayed for a randomly chosen song prediction. Participants were asked if a randomly chosen



feature in the given explanation contributes positively or negatively towards the systems prediction. For example, in figure 4, valence has a -14.06% weight towards dislike. Since the system predicts dislike, valence contributes negatively towards the decision class. For each framing condition, we measured how many participants answered the *vd* correctly (true/false). We did not use *vd* as a measure of overall understandability because it only tests understanding for one particular trial. A randomly chosen trial might not represent the entire condition. It did however provide a check for user involvement.

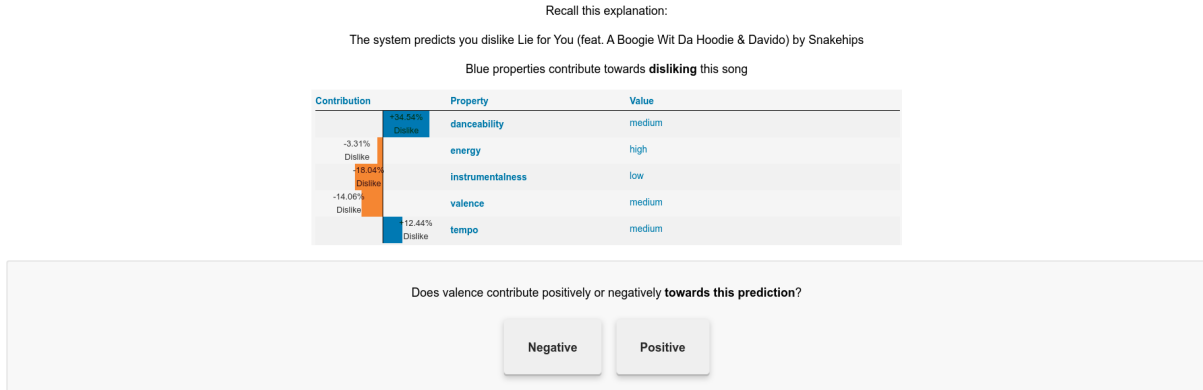


Figure 4: Valence Determination Task for the music AI system in the nucv with negative framing condition

### 3.5 Measured independent variables

We measured four other independent variables for exploratory analyses and to control for co-variations in our models. For our exploratory analysis, we measured *user belief* to investigate if explanations got influenced when participants agreed or disagreed with the systems prediction. As mentioned in section 2.4, user belief can affect how well participants understand our explanations. User belief was measured for each trial by asking the participants *Do you agree with this prediction? (No-Yes)*. Next, we measured *age* and *gender* since they could potentially co-vary with our duration measure. These were measured with demographic questions at the start of our study. Finally, to increase the accuracy of our duration measure, we measured and controlled how often people *opened* description boxes. As shown in our user interfaces, the description boxes contain feature description (eg fig 8). When people are trying to understand these feature descriptions, they might not be focused on our primary task.

### 3.6 Other measurements

Other measurements not extensively discussed in this report are mouse clicking behaviour for each trial (click positions and timestamps which could be used for process tracing), browser/platform meta data for each participant (for filtering mobile users), consent for online data storage and, emails from self-recruited/coursework participants.

## 4 Practical Implementation

In this section we outline our implementation of the music AI system and loan decision support system. We then describe the tools we used to conduct the study and our user interface. For both systems the procedure to generate explanations is outlined in figure 5.

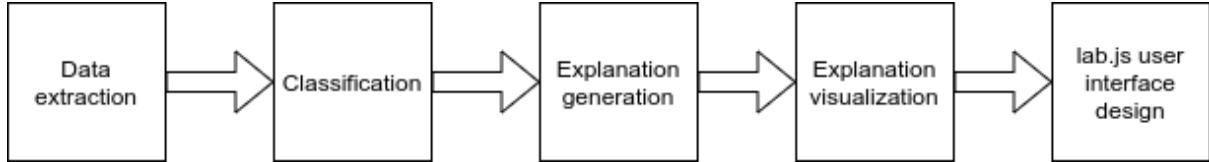


Figure 5: Implementation procedure of the explanations for the music AI system and loan decision support system

### 4.1 The music AI system

#### 4.1.1 Training data

Our dataset consists of 1997 songs extracted from Spotify using the Spotipy API<sup>1</sup>. Spotipy is a Python package based on the Spotify Web API<sup>2</sup>. Using Spotipy, for each song we extracted six audio features: danceability, energy, acousticness, instrumentalness, valence, and tempo. Since Spotify audio features provide detailed acoustic and psychoacoustic song properties, we used them in our music explanations. For each song, we classified each audio feature as high, medium and low based on mean feature values from our dataset. For our machine learning models, we used these Spotify audio features to predict the genre of each song. The genre of each song was based on the genre labels we assigned to them (Rock, Hip-hop, Pop or, Jazz). In order to assign songs with labels that are representative of their genres, we compiled our database using the most popular user created Spotify playlists for each genre. Spotify also provides genre labels but these are classified according to the songs artist so we did not use them. An artist might not necessarily represent the songs' genre.

#### 4.1.2 Machine Learning Model

In our study, participants chose their favourite genre (Rock, Hip-hop, Pop or, Jazz) and they received predictions that were suitable for their chosen genre. Based on the genre labels, our learner had to predict whether a user *"liked"* or *"disliked"* a particular song. Therefore, for each separate genre we labeled each song in our dataset as *"like"* if it belonged to the genre and *"dislike"* if it belonged to another genre. Using this binary labelling, we trained four different classifiers (one for each genre).

Our machine-learning models used a decision tree classifier with the scikit-learn package<sup>3</sup> in Python. A decision tree seemed most appropriate because our learner had to determine if a given feature with a particular value fits under the genre category based on multiple binary choices. We started off by splitting the dataset into a training set (70%) and test set (30%) for *each* genre label (so we had four classifiers). After training our learners using the entropy parameter, we obtained moderate in-sample accuracy's (rock = 78.8%, hip-hop = 76.8%, pop = 72.8% and jazz = 87.6%).

For the current study we aimed for a moderate, but not too high in-sample accuracy. Based on iterating with other datasets, we found that if a learner was 90-100% accurate, individual song explanations were unidirectional. Unidirectional explanations might not be very effective in our study because participants will not be able to reason whether individual feature explanations contribute negatively and positively for each song prediction. Since we wanted to test the differences between individual feature framing conditions, we wanted participants to reason in both directions for each trial. Therefore we needed less conclusive explanations that varied in different directions.

<sup>1</sup>Spotipy API

<sup>2</sup>Spotify Web API

<sup>3</sup>scikit-learn API

### 4.1.3 Explanation generation and visualization

For generating the explanations we used the LIME API <sup>4</sup> along with our self-built package, ArgueView and our visualizer ArgueView.js <sup>5</sup>. ArgueView along with its visualizer, was created by Sophia Hadash. It is a presentation tool that generates user friendly explanation output from explanation frameworks such as LIME and SHAP (Lundberg & Lee, 2017).

After training our four classifiers, we inputted each individual prediction in our test set into LIME to generate feature-maps. We then used the feature-maps along with feature descriptions for each audio feature, and inputted these into ArgueView to generate our explanations. Each ArgueView explanation contains the song prediction (like/dislike) (based on the prediction probabilities from LIME), the weights for each audio feature and its feature value (based on LIMEs feature map) and, the percentage of the final prediction that could not be explained by these features. For each song, this *unexplained* contribution is calculated as the difference between the sum of the feature weights and the highest prediction probability. The ArgueView explanation was then inputted into ArgueView.js to generate a visualization. An example visualization that we used for our music AI system can be seen in figure 6. Looking at this figure we see that contributions are shown as bar visualizations in a similar style to LIME. Using this method we generated multiple visualizations for each genre. To generate a fair representation of our learner, we randomly picked data-points for each genre. We did however filter for predictions that aligned with our labels since we did not want to show participants incorrect predictions. Although complete randomization would have been a more accurate representation, in the current study we are not testing the correctness of each prediction but rather, the framing of each explanation. Furthermore, to show only relevant feature contribution, we used a minimum threshold weight of 2% (based on iteration).

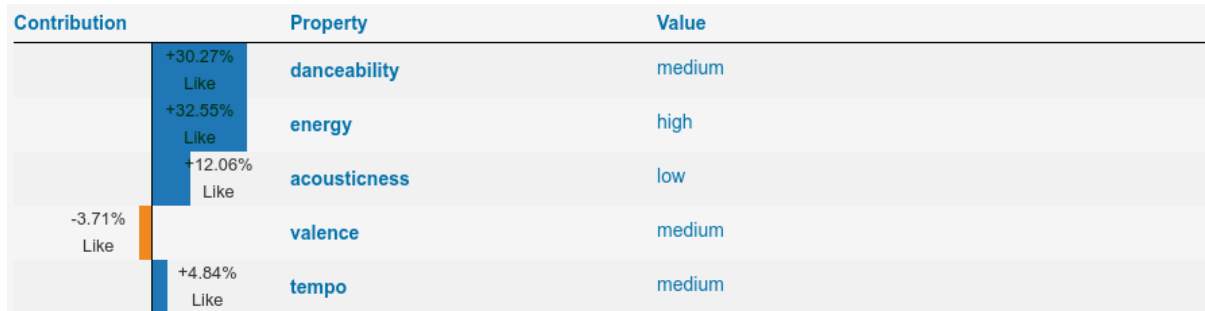


Figure 6: Visualization for a positive prediction (like) in the positive ucv & positive visualization framing condition (with blue contributing towards liking the song). Feature weights (Contribution), audio features (Property) and feature value (Value) are shown for each prediction.

## 4.2 The loan decision support system

### 4.2.1 Training data

For our loan system, we used the credit-g dataset in openML that classifies people into high and low credit risks based on 20 features (descriptions of each feature can be seen on the website) <sup>6</sup>. It contains 1000 datapoints collected by Dr.Frans Hoffmann in 1994. Although it is an old dataset, we felt that it was the most accessible and appropriate for our purpose.

### 4.2.2 Machine Learning Model

Similar to the music AI system we used sci-kit learn but this time, we used a random forest classifier which are essentially multiple decision trees. We used a random forest for this model because classifying based on 20 features would be more computationally intensive for a decision tree classifier. We first split the dataset into a training (80%) and test set (20%). Then we trained our classifier using the following parameters: number of trees = 250, criterion = entropy, max\_features = 0.45, max tree depth = 10, minimum samples to be classified as a leaf node = 6, and minimum samples for a split = 7. With these

<sup>4</sup>LIME API

<sup>5</sup>ArgueView & ArgueView.js

<sup>6</sup>Credit-g Dataset

parameters, our classifier achieved a 77% in sample accuracy which was moderately accuracy (like our Spotify models).

### 4.2.3 Explanation generation and visualization

We used LIME along with our ArgueView package and Argueview.js visualizer to generate our explanations. The procedure for generating the explanations was the same as our music AI system (see section 4.1.3). An example visualization of our loan decision support system is shown in figure 7.

Contribution	Property	Value
+5.75% Ineligible	checking_status	lesser than 0€
-9.06% Ineligible	savings_status	between 100€ and 500€
+3.46% Ineligible	other_parties	none
+5.16% Ineligible	property_magnitude	car

Figure 7: Visualization for a negative prediction (ineligible) in the negative ucv & negative visualization framing condition (with blue contributing towards loan ineligibility). Feature weights (Contribution), credit features (Property) and feature values (Value) are shown for each prediction.

## 4.3 User interface design

We implemented our user interface and our online experiment with the open source study builder, lab.js<sup>7</sup>. The loan (Figure 9) and music (Figure 8) system user interfaces both contained a prediction (extracted from ArgueView) along with an explanation, feature information and two questions (over two separate screens). In each explanation, visualization framing was manipulated with the phrases directly below each prediction and the color of the bars while ucv framing was manipulated with the text next to each feature bar. Below each explanation the unexplained contribution was displayed so participants can also judge how accurate the explanations are. To the right of the explanation, the participants could look up feature descriptions. By clicking on a description, participants got more information about how each feature is measured along with the categorization levels (for categorical features). For the music system, feature descriptions were extracted from the Spotify web API. For the loan system, they were extracted from the original dataset and slightly modified to clarify more technical loan terms. We used a "carousel design" so participants could open and close each feature description once they were done (and focus on the explanations). In the music system, below the unexplained contribution, participants could preview the predicted song. We directly downloaded song previews from the Spotify API so the previews don't have any region locking problems. Music previews were used to aid participants who might be unfamiliar with a predicted song. Below the user interface, participants were first asked our question measuring understandability (eg figure 8). On the next screen the same user interface was presented but this time they were asked whether they agreed with the systems prediction (user belief) (eg figure 9). We separated these questions because we wanted to measure the time it took participants to click on the first question (duration). Additionally, the answers boxes were large (for easy to click answers) and equally sized (to prevent bias).

<sup>7</sup>lab.js

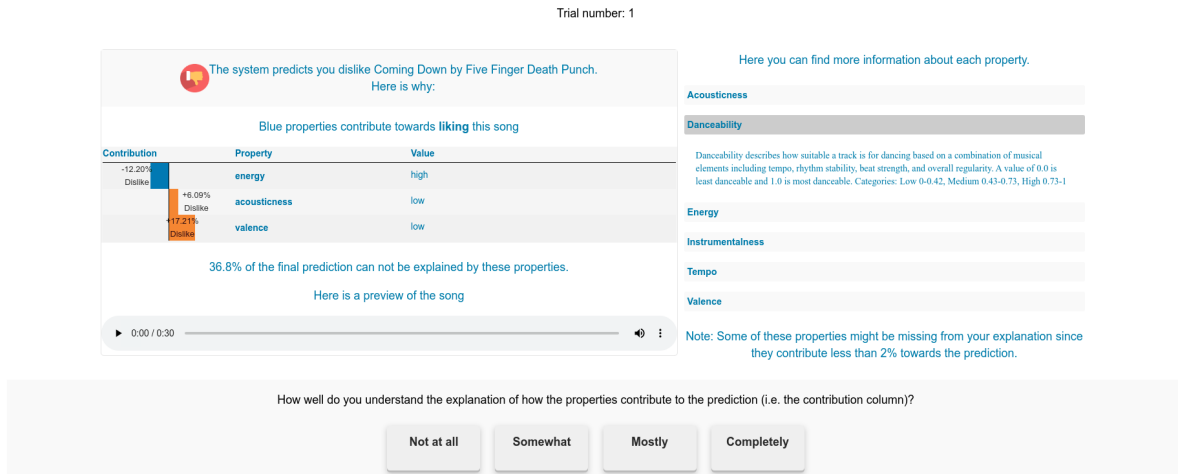


Figure 8: Music AI system with understandability question. The interface consists of the current trial number (top), prediction along with the explanation and song preview (middle left), feature descriptions (middle right), and a question (bottom).

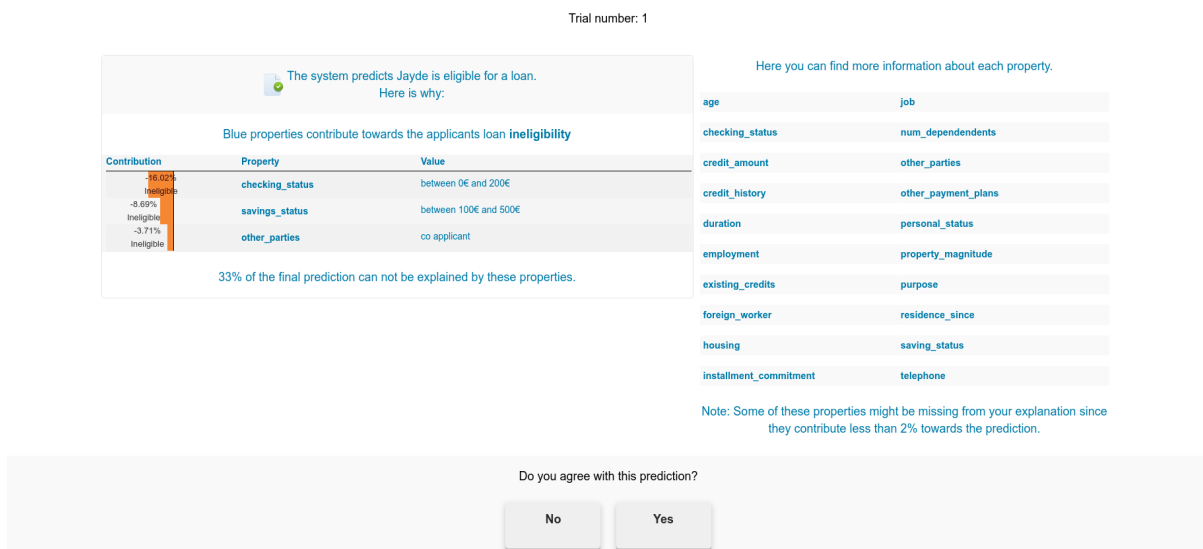


Figure 9: Loan decision support system with user belief question. The interface consists of the current trial number (top), prediction along with the explanation (middle left), feature descriptions (middle right), and a question (bottom).

## 5 Participants & Procedure

### 5.1 Participants

#### 5.1.1 Participant characteristics

Our study sample consists of 133 participants. Participants were sample from the JFS university database ( $n = 91$ ) a data science course (for more experienced users) and, based on convenience ( $n = 42$ ). Before recruiting from the university database, individuals who were above 70 years and had (corrected) poor eyesight were filtered out. More than half of the participants were aged between 18-24 ( $n = 77$ ) since majority of them were sample from our university database. Gender distribution between males ( $n = 64$ ) and females ( $n = 69$ ) was quite equal.

#### 5.1.2 Sampling procedures

Data was collected over a period of three weeks. The study was conducted online and hosted on our university server. An email was sent out with the study aim, duration, device usage, and compensation. Based on the pilot study we estimated the study to take 25 minutes. Since we measured reaction time, participants were requested to only do the study via a laptop or desktop PC. For participants sampled from the JFS university database, a lottery was organized with a 1-in-5 chance to win 25 euros with an alternative option to opt for course credits. Participants recruited from the data science course and via convenience sampling received a 5 euro gift card via email.

#### 5.1.3 Sample size determination

Using G\*Power, we conducted an a-priori power analyses on all three dependent measures while accounting for our between subject effect. In G\*power we used the: *ANOVA: Repeated measure, between factors power analysis*<sup>8</sup>. We tested our between subject manipulation, *UCV* because it has the largest effect on our power. Although there is no past literature on the effects of UCV on explanation understandability we do assume a medium effect size of interest. According to past literature in neuroscience and linguistics, framing does have a substantial effect on comprehension (e.g. Itkes & Mashal, 2016; Kuhlmann et al., 2016).

For *perceived understandability*, which is repeated 4 times, we assumed a moderate correlation between the repeated measurements ( $r = 0.5$ ) and a medium effect size of interest ( $f = 0.25$ ). We had three between subject groups (*nucv*, *no ucv* and *pucv*). Based on these inputs, we required 132 participants to achieve our power ( $\alpha = 0.05, 1 - \beta = 0.9$ ). For *understandability* and *duration* we had 24 repeated measures and assumed a slightly higher correlation between measurements since individual trials in each framing condition are more similar ( $r = 0.6$ ). We required 129 participants to show a medium effect size ( $f = 0.25$ ) with the same power ( $\alpha = 0.05, 1 - \beta = 0.9$ ). Our sample size determination was approved by our university review board. The final study sampled 133 participants so both requirements were met.

### 5.2 Pilot Study

Before deploying the study, a pilot study with seven participants was conducted. This was done in order to mitigate potential difficulties participants might face while going through the final study. Participants were self recruited and did not receive any compensation for the pilot. Also, to prevent biased results, they were not allowed to take part in the main study. Participants were invited to an online video call where they were provided with a link to the pilot study. They were then requested to share their desktop while going through the study. With screen sharing, researchers noted any observations related to clicking behaviour, the assigned UCV condition and the time it took to complete the study. To prevent any bias during the study, the researchers did not verbally intervene while the participants went through the study. However, if the participant had any questions they were free to ask. After completing the study, participants were informally interviewed about their thoughts on the clarity of the instruction pages and more on how they assessed the understandability score for each trial. After the interview, they were thanked for their participation. Apart from usability improvements, the pilot allowed us to improve our instruction screens and measurement of *understandability*. Initially our question stated: *How well do you understand this explanation?*. We noticed that each participants rated understandability based on different items in our user interface (e.g. based on the unexplained contribution, the feature values, the

<sup>8</sup>In our final model, we use multi level regression but this option is not available in G\*Power

feature descriptions) and the question needed to be more specific. Therefore, since we only wanted to investigate framing effects, we explicitly asked them if they understand how the visualizations contributed to their understandability.

### 5.3 Main Procedure

Participants were sent an invitation link via email or instant messaging. Once they opened the study, they were automatically assigned to a UCV condition. After agreeing to the consent form, participants were presented with a flowchart with an overview of the study. On the next screen, participants were randomly allocated to either interact with the music AI system or loan decision support system. If participants first received the music AI system, they chose their favourite genre out of Rock, Hip-hop, Rock and Jazz. They were then presented with instructions which included information about the AI, the music condition setup, and their task for each trial. Participants were required to carefully read this screen. On the next screen they were presented with a trial in a randomly allocated visualization framing condition. After the trials, participants answered a randomly chosen valence determination task (eg see figure 4) and a four item questionnaire measuring *perceived understandability*. Participants then received a warning that explanations were changing after which they completed the same procedure for the negative framing condition. Participants then interacted with the loan decision support system. Here they were immediately presented with an instruction screen that they were required to carefully read. Participants were asked to imagine a scenario where they were loan auditors. They were tasked with assessing the understandability of a newly implemented loan decision support system. After detailing the task, participants continued to follow the same procedure as the music AI system. Finally participants were thanked and told how they can receive compensation. For a visual overview of the procedure with all possible conditions, see figure 10. For an overview of our instruction screens see appendix A.



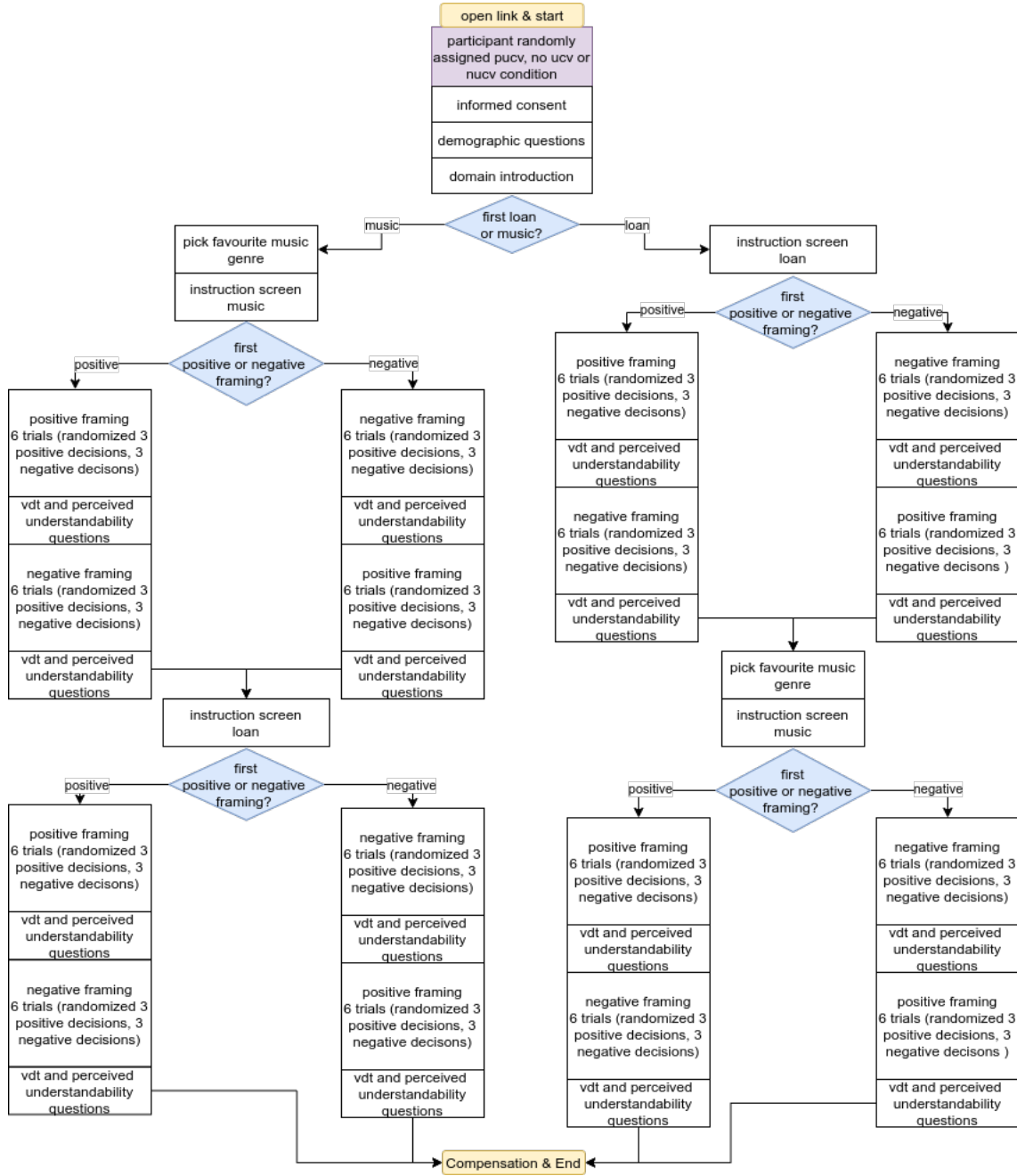


Figure 10: Participant study flowchart. *Note: Framing here refers to framing of the visualization. Understandability, duration and user belief were measured for each trial. vdt refers to the valence determination task. All choices (blue and purple) were made by the randomizer in lab.js so researchers did not know which participant was allocated to which condition.*

### 5.4 Data diagnostics

Before filtering our final dataset for analysis, a total on 139 participants completed the study. After analysing the time it took for participants to go through each instruction screen, five participants were excluded from receiving compensation. This is because they did not spend sufficient time reading the instructions (<20 seconds) and completed the study in less than half of the estimated study duration (< 12.5 minutes). Additionally, four of these participants used a mobile phone to complete the study which they were instructed not to do. From the 133 remaining participants, five more were excluded from the duration model since they used phones. However, since they spent sufficient time completing the study, they were included in the understandability and perceived understandability models.

Additionally some of the raw data was transformed. *Duration* and *opened* had skewed distributions



and were log transformed. *Perceived understandability* was calculated as the mean of the four items. For easier interpretability of our interaction effects, the categorical independent variables with higher order interactions were centered.

## 5.5 Analytic strategy

All analyses were conducted on *RStudio* using the *R* programming language. We initially examined the distribution and variance of our dependent measures and *vdt*. We also conducted a principle component analysis for our perceived understandability measure. We then correlated the dependent variables. We constructed three regression models (one for each measure of understandability). Since we have clustered data, we used multi-level regressions. Afterwards we constructed margins plots to interpret the direction of our main and interaction effects. We also checked each model for normality of residuals, homoskedasticity, and outliers.

For an overview of all the independent variables used in our models (and their definitions) see regression tables 6, 5, and 7.

## 6 Results

### 6.1 Descriptive statistics

Table 3 summarizes the distribution of the variables we used to measure understandability. It is evident that duration has a large spread. Although we log transformed duration, it might still be a noisy measurement. The most important measure is understandability, where participants were asked if they understood the explanation on each trials (with 3188 measures).

**Table 3** Descriptive statistics for each dependent variable measure

	mean	sd	min	max	n
duration (seconds)	30.55	41.75	0.036	704.25	3068
understandability	1.84	0.86	0	3	3188
perceived understandability	3.44	0.89	1	5	532

*Note: The duration timer started after each page had loaded.*

As an additional check to see if participants were engaged, for each condition we also administered a valence determination task. For all conditions, majority of the participants passed these tasks. For our most favourable condition according to our hypothesis (pucv with positively framed visualizations), 66.7% of the participants answered correctly. For the our least favorable (hypothesized) condition (nucv with negatively framed visualizations) 65% answered correctly.

### 6.2 Principle component analysis

For our perceived understandability measure, we conducted a principle component analysis. The output of the principle component analysis can be seen in table 4. The scree plot (figure 11) shows that more than 50% of the variance is explained by the first factor. Therefore, our perceived understandability items all fall under one dimension.

**Table 4** Principle component analysis for perceived understandability

Item	Factor 1	Factor 2	Factor 3	Factor 4
The explanations helped me get more insight into the given predictions.	0.44	0.72		0.43
The explanations felt clear to me.	0.56			
I felt that the explanations took me a lot of time to comprehend (reversed)	0.46			
The explanations were confusing to me (reversed)	0.52		0.7	0.4

*Note: Only factor loadings above 0.3 were included. Chronbachs  $\alpha = 0.73$  and Kaiser-Meyer-Olkin's measure of sampling adequacy = 0.71.*

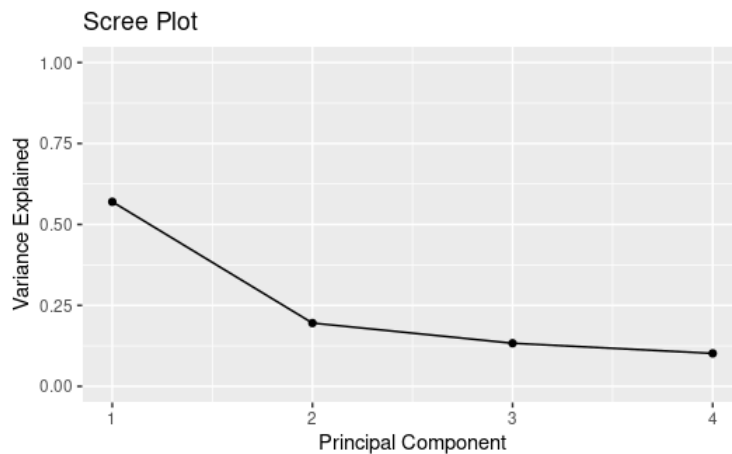


Figure 11: Scree plot depicting amount of variance explained by each factor

### 6.3 Correlations between dependent measures

Based on Pearson Product-Moment correlations, we found a moderate positive correlation between perceived understanding and understanding ( $r = 0.33$ ,  $p < 0.001$ ). Duration did not correlate with the other two measures.

### 6.4 Dealing with clustered data

For all three models, our data was clustered on the participant level. We find the largest clustering for perceived understandability ( $\rho = 0.44$ ) followed by understandability ( $\rho = 0.3$ ) and duration ( $\rho = 0.27$ ). The larger clustering is because perceived understandability is measured on the conditions level while the other variables are measured on the trial level. Therefore, it was necessary to use multi-level regression for all three models.

### 6.5 Assumption testing

We tested all three models for potential outliers, normality of residuals and homoskedasticity.

#### 6.5.1 Outliers

No outliers were noticeable in the understandability and perceived understandability models. In the duration model, we flagged five potential outlier points. Without the outliers, duration no longer violated the normality assumption and therefore these values were removed. The outliers were kept for the other two models. For a graph showing these outliers, see appendix B.

#### 6.5.2 Normality of residuals

A Shapiro-Wilk test for normality indicates that normality was not violated as all three measures have a W-score close to 1 ( $W_{understandability} = 0.98$ ,  $W_{perceived-understandability} = 0.97$ ,  $W_{duration} = 0.99$ ). For QQ-plots of the residuals, see appendix B

#### 6.5.3 Homoskedasticity

All three models violated the Breusch-Pagan test for homoskedasticity ( $p < 0.001$ ) and therefore, we used robust variants of the multi-level regression models. This did not make a large difference in the results.

### 6.6 Regression models

To test our hypotheses and conduct our exploratory analyses, we use three regression models. Recall that understandability is our main dependent variable that measures how well participants understand each trial. Duration measures the time it takes for participants to respond to the understandability question. In this context it complements the more subjective measurement for understandability as both measurements are taken on the trial level. Finally perceived understandability measures how participants feel about each condition after going through all six trials for a given condition.

After testing all assumptions we used robust multi-level regression to model our dependent variables. Tables 5, 6 and 7 show the understandability<sup>9</sup>, duration and, perceived understandability models respectively.

Some main effect variables such as gender and age were included as covariates. For example in table 5 we see that older people (55+) measured lower on understandability. In table 6 we also see that younger participants and males measured higher of duration. Also, from table 6 we see that when participants were more focused on feature descriptions, they measured lower on duration. Finally, from table 5 we see participants measure lower on understandability when a larger percentage of the systems predictions could not be explained by the underlying features. These effects won't be discussed in detail since it is not the main focus of our research. The other variables that were used to answer our research questions are elaborated on in the upcoming sections. Definitions of all the variables can be found in the regression tables.

<sup>9</sup>Since understandability was an ordinal variable with four levels, we also analysed a multi level ordinal logistic regression and found the same results. To maintain ease of interpretation, we only show the multi-level regression results for this report.

**Table 5** Results of the robust multi-level regression model with understandability as dependent variable and participant as random intercept.

	<b>B</b>	<b>SE</b>	<b>t-value</b>	<b>p</b>
n = 133				
intercept	1.89	0.049	38.89	***
Hypotheses				
framing <sub>ucv</sub> (nucv = -1, no ucv = 0, pucv = 1)	0.17	0.037	4.73	***
framing <sub>visualization</sub> (1 = positive)	0.045	0.012	3.65	***
ucv (nucv, pucv = -1, no ucv = 2)	0.023	0.022	1.03	0.31
decision (like/eligible = 1)	0.0043	0.0168	0.26	0.79
framing <sub>visualization</sub> * decision	0.0071	0.012	0.58	0.56
Exploratory				
first impression (trial 1 = -5, trial 2-6 = 1)	0.019	0.005	3.48	***
domain (music = 1)	-0.067	0.012	-5.36	***
user belief (agree = 1)	0.034	0.015	2.32	*
first impression * domain	-0.015	0.0055	-2.64	**
framing <sub>ucv</sub> * framing <sub>visualization</sub>	0.053	0.015	3.57	***
domain * framing <sub>visualization</sub>	0.027	0.012	-2.19	*
framing <sub>ucv</sub> * framing <sub>visualization</sub> * user belief	-0.058	0.017	-3.32	***
Covariates				
unexplained	-0.0022	0.00075	-2.92	**
age (25-34)	-0.011	0.075	-0.15	0.88
age (35-44)	0.067	0.15	0.44	0.66
age (45-54)	-0.055	0.12	-0.44	0.66
age (55-64)	-0.31	0.13	-2.35	*
age (65+)	-0.58	0.21	-2.79	**
framing <sub>visualization</sub> * user belief	0.027	0.014	1.82	0.067
framing <sub>ucv</sub> * user belief	-0.018	0.018	-0.99	0.32
framing <sub>visualization</sub> * ucv	0.076	0.0089	8.59	***
domain * ucv	-0.012	0.0089	-1.33	0.18
domain * ucv	-0.015	0.015	-0.95	0.34
domain * framing <sub>visualization</sub> * ucv	-0.021	0.0088	-2.43	*
domain * ucv * framing <sub>visualization</sub>	-0.0210	0.015	-1.41	0.15

\*\*\*  $p < 0.001$ , \*\*  $p < 0.01$ , \*  $p < 0.05$ .

*Hypotheses variables test our three hypothesis while exploratory variables test the effects of first impressions, domain, user belief and their interactions with our manipulations. Other variables were included to control for co-variations.*

*Variables: framing<sub>ucv</sub>: UCV framed positive (1) versus UCV framed negative (-1). No UCV is 0; framing<sub>visualization</sub>: visualization framed positive (1) versus negative (-1); ucv: The inclusion (2) or exclusion of UCVs (pucv=-1, nucv=-1). UCV was recoded this way to test H1a; first impression: The first trial (-5) for a condition versus trials 2-6 (1); unexplained: the percentage of a decision that can't be explained by the features; user belief: whether the user agrees (1) or disagrees (-1) with the predictions; domain: music AI system (1) or the loan decision support system (-1); age: participants age*

**Table 6** Results of the robust multi-level regression model with log(duration) as dependent variable and participant as random intercept.

n = 128	<b>B</b>	<b>SE B</b>	<b>t-value</b>	<b>p</b>
intercept	2.66	0.046	57.4	***
<b>Hypotheses</b>				
ucv (pucv,nucv = -1, no ucv = 2)	-0.0062	0.018	-0.33	0.74
framing <sub>ucv</sub> (nucv = -1, no ucv = 0, pucv = 1)	-0.031	0.031	-0.99	0.32
framing <sub>visualization</sub> (positive = 1)	-0.022	0.01	-2.14	*
decision (like/eligible = 1)	0.029	0.014	2.04	*
framing <sub>visualization</sub> * decision	-0.028	0.01	-2.73	**
<b>Exploratory</b>				
domain (music = 1)	0.22	0.012	18.38	***
user belief (agree = 1)	-.032	0.012	-2.59	**
first impression (trial 1 = -5, trials 2-6 = 1)	-0.11	0.0065	-18.04	***
framing <sub>ucv</sub> * framing <sub>visualization</sub>	-0.015	0.012	-1.21	0.23
framing <sub>ucv</sub> * user belief	-0.012	0.015	-0.81	0.42
framing <sub>visualization</sub> * user belief	-0.031	0.012	-2.57	**
domain * framing <sub>visualization</sub>	0.046	0.01	4.36	***
first impression * domain	0.034	0.0064	5.24	***
framing <sub>ucv</sub> * framing <sub>visualization</sub> * user belief	-0.019	0.015	-1.31	0.19
<b>Covariates</b>				
correct vdt (true = 1)	-0.043	0.026	-1.69	0.091
unexplained	0.0026	0.00064	4.09	***
gender (male = 1)	0.18	0.054	3.37	***
age (25-34)	-0.099	0.063	-0.16	0.87
age (35-44)	-0.054	0.143	-0.40	0.69
age (45-54)	0.15	0.11	1.43	0.15
age (55-64)	0.75	0.12	6.45	***
age (65+)	-0.078	0.16	-0.47	0.64
opened	0.32	0.01	29.12	***
opened * first impression	0.012	0.0031	3.90	***
opened * domain	-0.053	0.011	-5.04	***
domain * ucv	-0.0015	0.0074	-0.2	0.84
framing <sub>visualization</sub> * ucv	-0.006	0.0075	-0.81	0.42
opened * first impression * domain	-0.0099	0.003	3.24	**
domain * framing <sub>visualization</sub> * ucv	0.024	0.0074	3.22	**
framing <sub>ucv</sub> * framing <sub>visualization</sub> * decision	-0.019	0.015	-1.31	0.57

\*\*\*  $p < 0.001$ , \*\*  $p < 0.01$ , \*  $p < 0.05$ .

*Hypotheses variables test our three hypothesis while exploratory variables test the effects of first impressions, domain, user belief and their interactions with our manipulations. Other variables were included to control for co-variations.*

*Variables: framing<sub>ucv</sub>: UCV framed positive (1) versus UCV framed negative (-1). No UCV is 0; framing<sub>visualization</sub>: visualization framed positive (1) versus negative (-1); ucv: The inclusion (2) or exclusion of UCVs (pucv=-1, nucv=-1) in the explanation. UCV was recoded this way to test H1a; opened: number of times a description box is opened (log transformed); first impression: The first trial (-5) for a condition versus trials 2-6 (1); user belief: whether the user agrees (1) or disagrees (-1) with the predictions; unexplained: the percentage of a decision that can't be explained by the features; domain: music AI system (1) or the loan decision support system (-1); correct vdt: whether participants correctly answered the valence determination task (1 = true, -1 = false); age: participants age; gender: participants gender*

**Table 7** Results of the robust multi-level regression model with perceived understandability as dependent variable and participant as random intercept.

	<b>B</b>	<b>SE B</b>	<b>t-value</b>	<b>p</b>
n = 133				
intercept	3.48	0.054	64.26	***
Hypotheses				
framing <sub>ucv</sub> (nucv = -1, no ucv = 0, pucv = 1)	0.036	0.044	0.82	0.41
framing <sub>visualization</sub> (1 = positive)	0.1	0.02	5.04	***
ucv (nucv,pucv = -1, no ucv = 2)	0.0082	0.023	0.35	0.72
decision (like/eligible = 1)	.00026	0.011	-0.23	0.82
framing <sub>visualization</sub> * decision	0.0025	0.012	0.22	0.83
Exploratory				
domain (music = 1)	0.022	0.012	1.96	0.051
framing <sub>ucv</sub> * framing <sub>visualization</sub>	0.13	0.024	5.21	***
framing <sub>visualization</sub> * domain	-0.018	0.011	-1.52	0.12
user belief (agree = 1)	0.062	0.014	4.48	***
framing <sub>ucv</sub> * framing <sub>visualization</sub> * user belief	-0.020	0.016	-1.26	0.21
Covariates				
correct vdt (true = 1)	0.0075	0.029	0.26	0.8
ucv * domain	0.0072	0.0082	0.87	0.38
framing <sub>ucv</sub> * correct vdt	0.049	0.034	1.48	0.14
framing <sub>visualization</sub> * correct vdt	-0.023	0.026	-0.89	0.37
framing <sub>visualization</sub> * ucv * domain	-0.053	0.0083	-6.42	***
framing <sub>visualization</sub> * user belief	0.012	0.013	0.90	0.37
framing <sub>ucv</sub> * user belief	0.035	0.017	2.11	*
framing <sub>ucv</sub> * framing <sub>visualization</sub> * correct vdt	-0.14	0.031	-4.37	***

\*\*\*  $p < 0.001$ , \*\*  $p < 0.01$ , \*  $p < 0.05$ .

*Hypotheses variables test our three hypothesis while exploratory variables test the effects of domain and interaction effects with our manipulations. Other variables were included to control for co-variations.*

*Variables: framing<sub>ucv</sub>: UCV framed positive (1) versus UCV framed negative (-1). No UCV is 0; framing<sub>visualization</sub>: visualization framed positive (1) versus negative (-1); ucv: The inclusion (2) or exclusion of UCVs (pucv=-1, nucv=-1) in the explanation. UCV was recoded this way to test H1a; domain: music AI system or the loan decision support system; user belief: whether the user agrees (1) or disagrees (-1) with the predictions*

Before continuing the analysis, recall our research questions:

**RQ1:** How do feature explanations containing underlying continuous variables influence the ability to understand explanations?

**RQ2:** How does a framed visualization with feature explanations and its valence alignment with the final prediction influence the ability to understand recommendations?

The next sections attempt to answer these questions by testing our hypotheses and interaction effects.

## 6.7 Hypotheses testing

### 6.7.1 Underlying continuous variable effects

The first research question investigates how UCVs influence the understandability of our explanations. Recall that we formulated two hypotheses:

**H1a:** Feature explanations with underlying continuous variables will be easier to understand than explanations without underlying continuous variables.

**H1b:** Feature explanations with positive underlying continuous variables will be easier to understand than explanations with negative underlying continuous variables.

To test H1a we look at the main effect of the *ucv* variable where UCV is compared to no UCV. In tables 5, 6 and, 7 we see that this variable is not significant and therefore H1a is rejected. However, *ucv* does

not distinguish between positively framed UCVs and negatively framed UCVs.

To compare the UCV framing condition and test H1b, we look at the variable  $framing_{ucv}$ . We measure a better understanding for positively framed UCVs compared to negatively framed UCVs ( $B = 0.17$ ,  $t = 4.73$ ,  $p < 0.001$ ) but only for the understandability model. Figure 12 shows marginal plots of  $framing_{ucv}$  for all three models. The understandability plot (top right), shows that pucv is easiest to understand followed by no ucv and finally nucv. In understandability scores, notice that nucv and pucv are roughly the same distance away from no ucv. In our  $ucv$  variable, this might indicate that pucv and nucv are cancelling each other out possibly indicating that the  $ucv$  variable does not control for framing effects in UCV. Looking at the duration (top left) and perceived understandability (bottom left) plot, we see that the effect seems to be in the same direction. However the error bars in both these plots are larger than the understandability plot. In summary, the results for our first hypothesis indicate that positive UCVs are easier to understand than negative UCVs. For localized XAI tools, this implies that if UCVs are added to an explanation, it might be more beneficial to frame them positively.

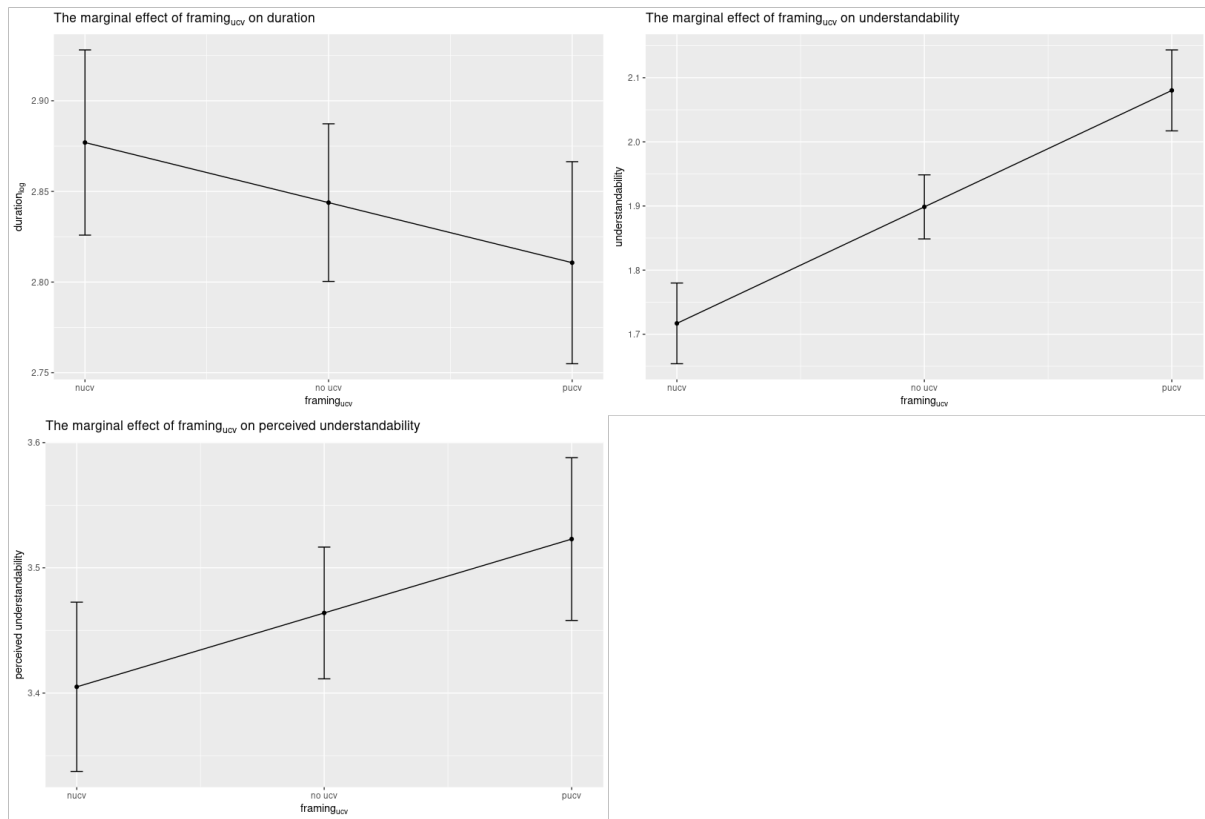


Figure 12: Margin plots with the main effect of framed ucvs on duration (top left where lower values measure quicker understanding), understandability (top right), and perceived understandability (bottom left)

### 6.7.2 Visual framing effects

The second research question investigates how visualization framing and its alignment with the prediction influences understandability. To examine visualization framing we look at H2:

**H2:** Feature explanations with positively framed visualizations will be easier to understand than explanations with negatively framed visualizations.

We examine the main effect of framed visualizations by looking at the  $framing_{visualization}$  variable in tables 5, 6 and, 7. Figure 13 shows the margin plots for all three models.

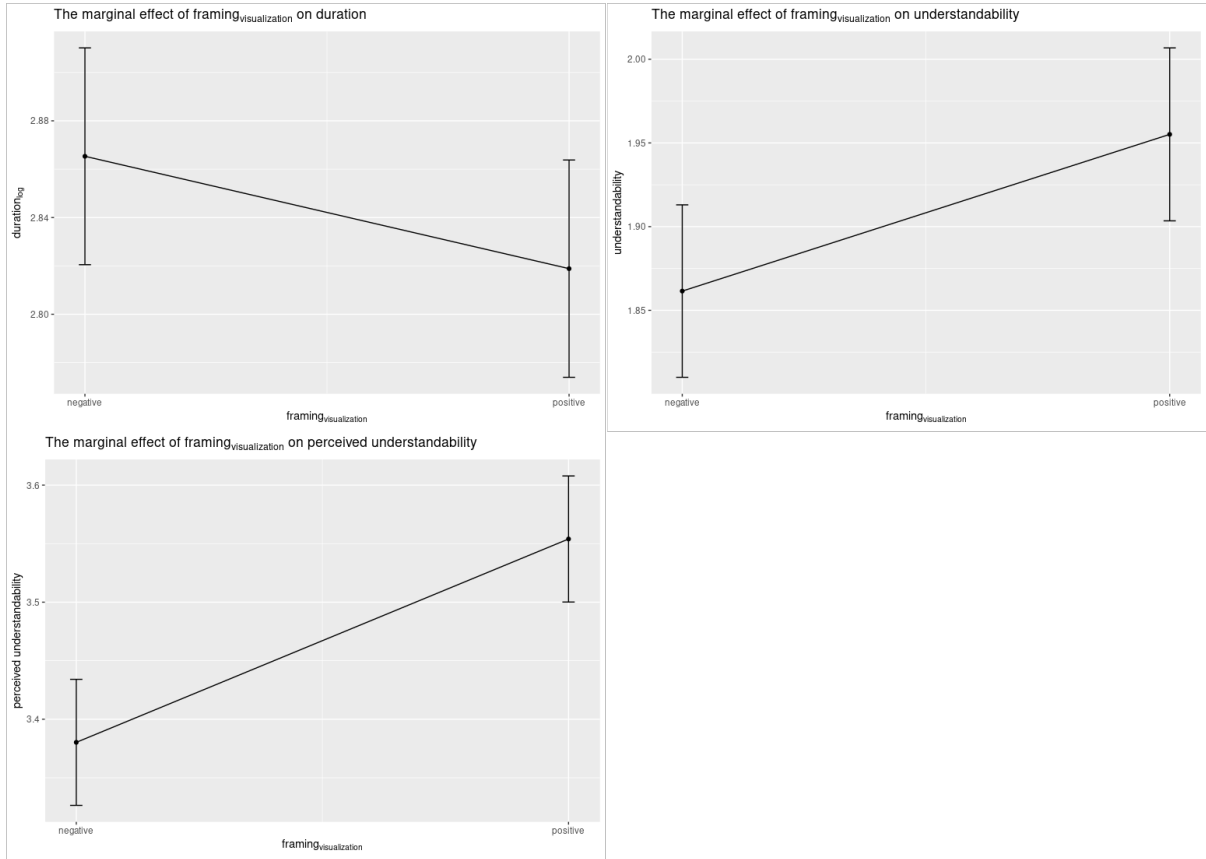


Figure 13: Margin plots with the main effect of visualization framing on duration (top left, Lower duration implies faster understandability), understandability (top right), and perceived understandability (bottom left).

We measured a better understanding for positively framed visualizations compared to negative framed visualizations for duration ( $B = -0.022$ ,  $t = -2.12$ ,  $p < 0.01$ ), understandability ( $B = 0.045$ ,  $t = 3.64$ ,  $p < 0.001$ ) and perceived understandability ( $B = 0.1$ ,  $t = 5.04$ ,  $p < 0.001$ ). Therefore H2 is supported by all three models. This might imply that localized XAI tools are comparatively less understandable because they currently also use negatively framed visualizations (e.g. figure 1).

### 6.7.3 Alignment between framing and prediction

To answer our second research question, we also examined the interaction effect between framing and prediction. H3 states the following:

**H3:** Feature explanations where the visualization framing and prediction are aligned will be easier to understand than explanations where the visualization framing and prediction are misaligned.

In the duration model, we find that the main effect of *framing<sub>visualization</sub>* is different for different predictions ( $B = -0.028$ ,  $t = -2.73$ ,  $p < 0.01$ ) however we don't find this for the understandability and perceived understandability models. To understand this interaction in more detail see figure 14.



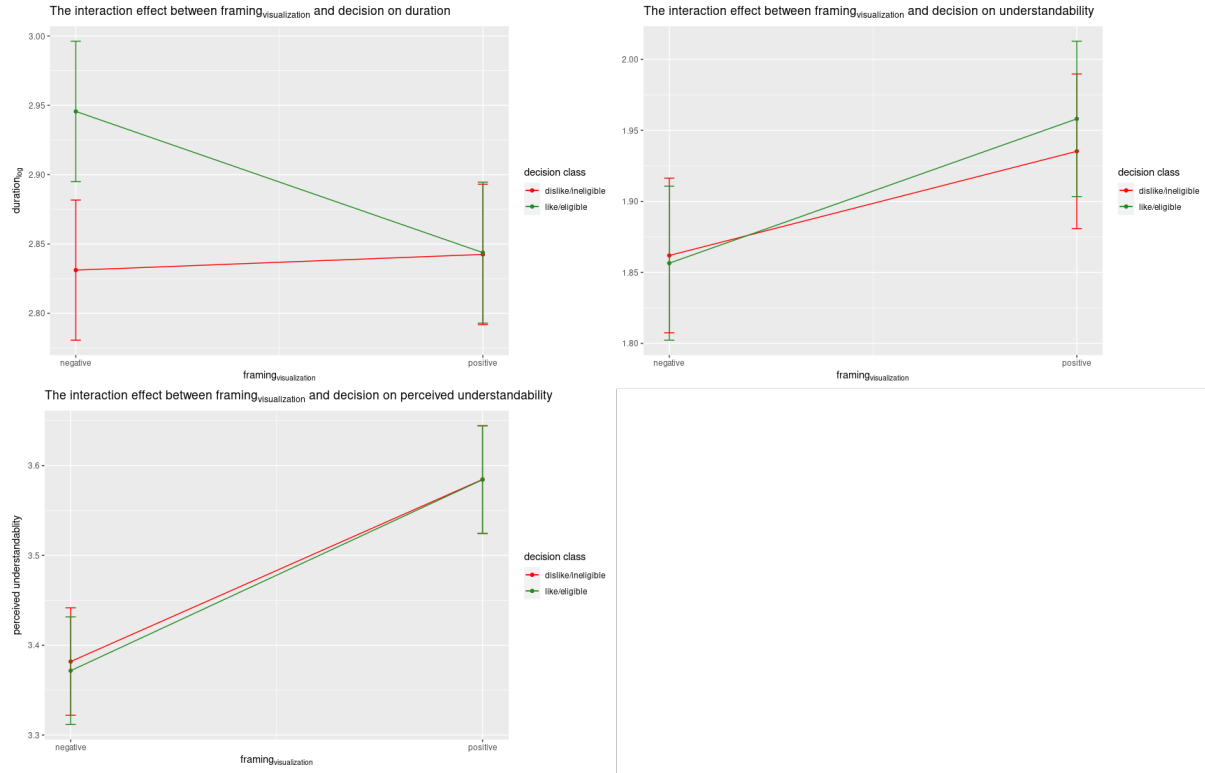


Figure 14: Margins plot of the interaction effect between visualization framing and prediction on duration (top left where a lower duration value implies faster understandability), understandability (top right) and, perceived understandability (bottom left) .

In the duration model, we find that for a positive decision class, it takes longer time to understand negatively framed visualizations compared to positively framed visualizations. We don't find the opposite effect for the negative decision class. Therefore, we find partial evidence for H3 in our duration model. Furthermore, looking at the understandability and perceived understandability plots, we do see that positively framed visualizations seems to be easier to understand than negative framed visualizations, regardless of decision class. This finding aligns with our second hypothesis. Both understandability plots emphasize that positively framed visualization might be more understandable for localized XAI tools.

## 6.8 Interactions effects

### 6.8.1 UCV and visualization framing

We find that the main effect of  $framing_{ucv}$  is different for different levels of  $framing_{visualizations}$  in the understandability ( $B = 0.053$ ,  $t = 3.57$ ,  $p < 0.001$ ) and the perceived understandability ( $B = 0.13$ ,  $t = 5.21$ ,  $p < 0.001$ ) models, but not for the duration model. Figure 15 provides us with more details on the direction of the interaction effect.

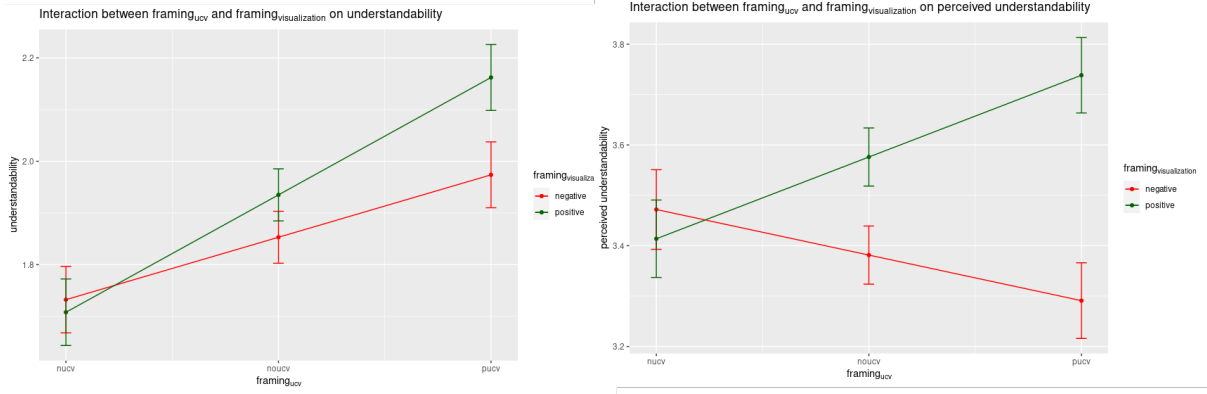


Figure 15: Margins plot of the interaction effect between ucv and framing on understandability (left) and perceived understandability (right)

From the understandability plot (left), it is clear that positively framed UCVs along with positive explanations are most understandable. Our perceived understandability plot (right) shows that even after all six trials, this condition is most understandable. This implies that we can make local XAI tools more understandable with some modifications. To improve understandability, explanations must not only be visualized with respect to the positive decision class, but each feature must contain a label that is framed with respect to the positive decision class. For example, in the LIME bar visualization from the introduction in figure 1, positive contributions like *credit history* can contain a label such as *+good*. Negative contributions, like checking status can contain a positive label such as *-good*. Coming back to figure 15, we also see that positively framed visualizations with no ucvs still measure quite high on understandability and perceived understandability. Therefore it might be sufficient to simply frame local XAI visualizations positively. For example, in figure 1, instead of the *good* and *bad* labels above the bar visualization, it would be better to frame it as *blue contributes towards good*. Of course, this is not the only way that visualizations can be modified in local XAI tools but our results indicate that framing visualizations positively does increase understandability. In figure 15 we do see that negatively framed visualizations measured the lowest on understandability. This enforces the point that negatively framed visualizations in local XAI tools are not very understandable. In fact, after six trials we see that negatively framed visualization with positively framed UCVs were the least understandable. Therefore, if negatively framed visualizations are still going to be used in local XAI models, they should not be combined with positively framed UCVs.

### 6.8.2 UCV, framing and, user belief

After each trial, participants were asked if they agreed or disagreed with a prediction. If participants disagree, we wanted to see if explanations have different understandability scores due to potential confirmation bias (Kahneman & Tversky, 1972). We measure this in tables 5, 6 and, 7 with the variable *user belief*. When user belief was aligned (agree), we measured shorter durations ( $B = -0.032$ ,  $t = -2.59$ ,  $p < 0.01$ ), higher understandability scores ( $B = 0.034$ ,  $t = 2.32$ ,  $p < 0.05$ ) and, higher perceived understandability scores ( $B = 0.062$ ,  $t = 4.48$ ,  $p < 0.001$ ) than when user belief was misaligned (disagree). This already implies that user belief does impact how well participants understood our explanations. In table 5 we look at the three way interaction to see how user belief particularly influences our framing manipulations; framing<sub>ucv</sub> and framing<sub>visualization</sub>. We not only find that the main effect of framing<sub>ucv</sub> is different for different levels framing<sub>visualization</sub> but also for different levels of user belief ( $B = -0.058$ ,  $t = -3.32$ ,  $p < 0.001$ ). The three way interaction effect was not found for the duration and perceived understandability models. To understand the directionality of this three way interaction see figure 16.

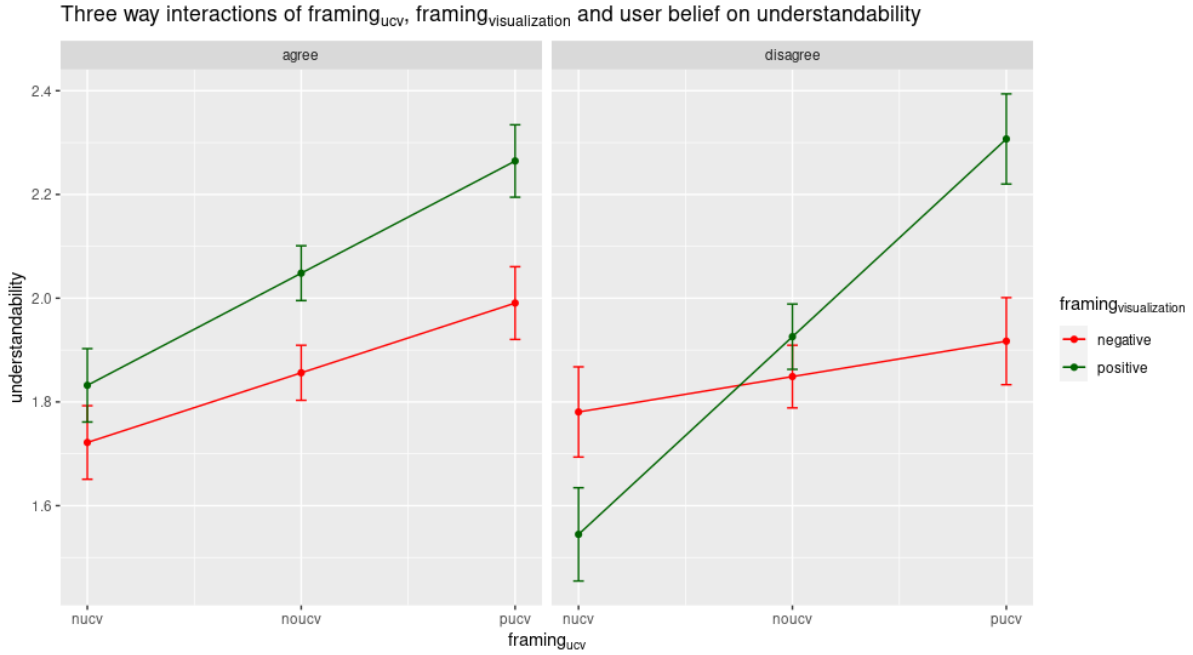


Figure 16: Margins plot of the three way interaction effect between  $\text{framing}_{ucv}$ ,  $\text{framing}_{visualization}$  and user belief (left = agree/aligned, right = disagree/misaligned) on understandability

For both plots in figure 16 we see a diminished effect of UCV when visualizations are framed negatively. The diminished effect is similar to the understandability plot in figure 15. When participants disagree with a prediction however, we do see that the effect of UCV on understandability is larger, for positively framed visualizations than when participants agree. Comparing both plot in figure 16 there definitely seems to be some level of confirmation bias with framed explanations. In fact, once again, this is in line with the conclusion from our first hypothesis that it is important to frame UCVs positively. When user belief is misaligned with the systems’ prediction and the visualization is framed positively, it is even more important to frame UCVs positively rather than negatively. For negatively framed visualizations, user belief does not seem to matter too much.

### 6.8.3 UCV, framing and, domain

To see if our results generalize across domains, we tested all our explanations on a music AI system and a loan decision support system. We expected that our music AI system would be harder to understand than our loan decision support because music predictions contain more unquantifiable bias (Doshi-Velez & Kim, 2017). The main effect of *domain* shows that the loan decision support system takes participants lesser time to understand ( $B = 0.22$ ,  $t = 18.38$ ,  $p < 0.001$ ) and is more understandable ( $B = -0.067$ ,  $t = -5.36$ ,  $p < 0.001$ ) than the music AI system. The same effect was not found for the perceived understandability model. Our  $\text{framing}_{ucv}$  variable did not change across domains. However, we found that the main effect of domain is different for different visualization framing conditions for the duration ( $B = 0.046$ ,  $t = 4.36$ ,  $p < 0.001$ ) and understandability ( $B = 0.027$ ,  $t = -2.19$ ,  $p < 0.05$ ) models. We did not find this interaction effect for perceived understandability. Figure 17 shows the interactions effects.

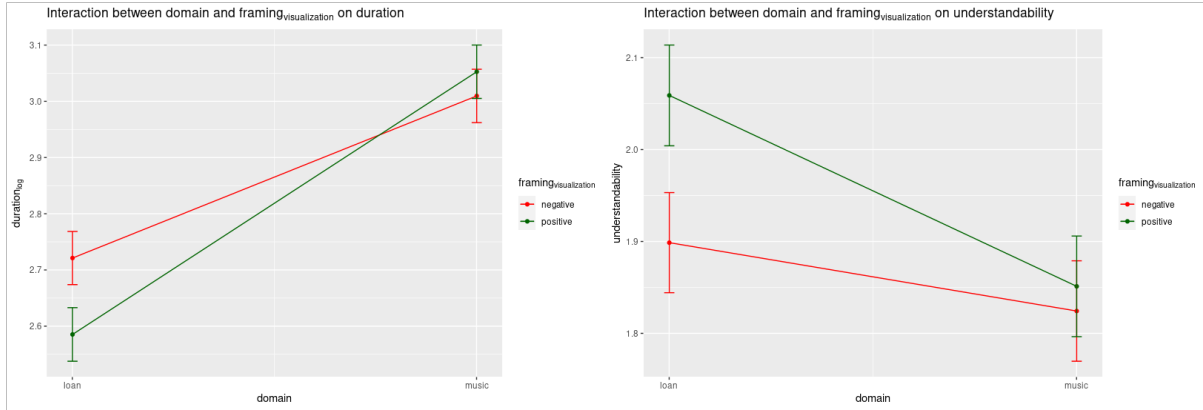


Figure 17: Margins plot of the interaction between domain and framing<sub>visualization</sub> on duration (left where lower is better understanding) and understandability (right)

From the understandability plot (left) we once again see that our loan system is generally easier to understand than our music system. In particular, we see that for the loan system, positively framed explanations are quicker and easier to understand than negatively framed explanations. Visualization framing conditions don't seem to matter for the music system (given the error bars).

To investigate domain effects even further, we looked at how the first trial for each condition compared with the remaining five trials. Since participants had to familiarize themselves with the user interface, the explanations in our first trial might have been harder to understand. In tables 5 and 6 we added the *first impression* variable coded as -5 for the first trial, and 1 for the other five trials (to ensure balanced weights). Measuring duration, we found that the first trial took more time to understand compared to subsequent trials ( $B = -0.11$ ,  $t = 18.04$ ,  $p < 0.001$ ). Measuring understandability, we also found that participants had a harder time understanding the first trial in comparison to subsequent trials ( $B = 0.019$ ,  $t = 3.48$ ,  $p < 0.001$ ). Across domains, we found that the main effect of *first impression* is different for different domains in both, the duration ( $B = 0.034$ ,  $t = 5.24$ ,  $p < 0.001$ ) and understandability ( $B = -0.015$ ,  $t = -2.64$ ,  $p < 0.001$ ) models. Figure 18 provides more information on these interaction effect. We do see that in both models there is no difference across domains for the first trial. So when an explanation is initially presented, it seems that these explanations might generalize across domains. However for the remaining five trials we see that the loan decision support system is more understandable than the music AI system. This might imply that explanations should be presented sparingly. Although this is not conclusive evidence (since we don't test this null effect), it does align with XAI practices stating that explanations should only be presented when the context calls for it (Miller, 2019).

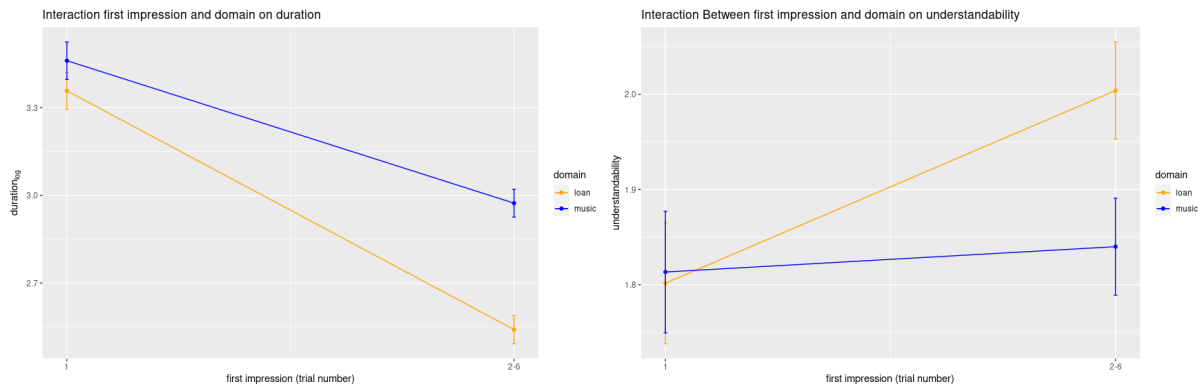


Figure 18: Margins plot of the interaction between first impression and domain on duration (left where lower is better understanding) and understandability (right)

## 7 Discussion

This section provides a discussion of the results. We start off by discussing how framed UCVs and visualizations influence explanation understandability. We then discuss how the understandability of such explanations could be influenced by user belief, application domain and learning effects over multiple trials. We then discuss the limitations of the current study. Finally we discuss the relevance of our study and future research directions.

### 7.1 The effect of framed UCVs and visualizations on user understandability

Our initial evidence suggests that the addition of an underlying continuous variable (UCV) does not influence understandability. But this comparison between UCV and no UCV does not distinguish between positively and negatively framed UCVs. In fact, we found that framing a UCV to be positive or negative makes a significant difference. Features with positively framed UCVs are the most understandable followed by no UCVs. Explanations with negatively framed UCVs are the least understandable. These results support our first hypothesis and imply that the usage of a UCV improves understandability, but only when it is framed positively. Furthermore, our results indicate that users find explanations more understandable when the visualizations are framed with respect to the positive decision class. This provides support for our second hypothesis. Looking back at previous findings, both of these results are in line with the Polyanna Principle which states that positive information is easier to process than negative information (Boucher & Osgood, 1969; Matlin, 2016). It is also in line with the studies by Itkes and Mashal (2016) who found that people have more difficulty understanding phrases where the overall valence of a words pair is negative (eg. 'Love Triangle'). Although there was not much research on framing in XAI research, the results are in line with the observations found by Stumpf et al. (2007). They noticed that negative keyword-based explanations confused participants while tasked with understanding the decisions of an email classifier.

Out of all framing conditions, we found that a positively framed UCV for each feature along with a positively framed visualization, was the most comprehensible. We also found that positively framed visualizations were easier to understand for both, explanations containing positively framed UCVs and, for explanations where only the visualization framing was manipulated (i.e. no UCV was present). The same effects were also found in our perceived understandability model. Unlike the results of Cramer et al. (2008) however, our perceived understandability measure was positively correlated with our understandability measure implying that participants did in fact understand the explanations as well as they thought. We also see this because majority of the participants passed the valence determination tasks. However, since 35% failed the valence determination task in all conditions, it does show that participants found these questions tricky. In contrast to our "positive" results, negative UCVs with negatively framed visualizations had the least understandable explanations. This result is in line with psycholinguistic research showing that comprehension decreases when sentences contain two or more negations (Corblin, 1996; Sherman, 1976). It is also in line with Spinoza's theory that people are required to pay more attention when words are framed negatively, making them hard to understand (Gilbert et al., 1990). Our results also imply that if UCV and visualization framing are aligned, people understand the explanations well when they are both positively framed. When they are both negatively framed explanation are harder to understand. In fact, positive UCVs with negatively framed visualizations were perceived to be the least comprehensible condition. This "misaligned" effect is in line with the study by Kuhlmann et al. (2016) who also found that German misaligned bivalent words had the lowest comprehension (eg "Erotikengel" (erotic angel) vs "Diplomübelkeit" (diploma nausea)).

Finally, we found partial evidence for our third hypothesis. When the valence of the decision class was aligned with the framing of the visualization, it took lesser time to understand than when these were misaligned. This is in line with psycho-linguistic literature stating that words valences that are aligned with the overall truth value of a statement are more comprehensible (Carpenter & Just, 1975; Clark & Chase, 1972). However, our duration model is a noisy measure compare to our understandability model. For both understandability models, we did not find evidence for this hypothesis. In fact, we found that positively framed visualizations are more understandable, regardless of decision class. Since this also aligns with our second hypothesis, we think that it might be better to frame visualizations positively. Perhaps psycho-linguistic literature does not translate well in this case because a decision class can't be likened to the truth value of a statement. This could be because a prediction made by the system does not imply that it is a fact. Even the users' perceived ground truth (that they agree/disagree with the prediction) might not align with a system's prediction.

Our results also have implications in the XAI domain. When a user receives an explanation, they want to understand the semantics behind why a certain prediction was made [Itkes and Mashal \(2016\)](#). In order to communicate the explanations clearly, explanation must not contain any obscurities ([Miller, 2019](#)). However, in XAI interfaces that frame explanations with respect to decision class labels, they might be creating obscurities without their knowledge. Obscurities such as double negations (*credit history contributes negatively towards not getting a loan*) are much more likely to occur in negative explanations. Therefore, negative explanations might dissuade users from using local explanation tools such as LIME ([Ribeiro et al., 2016](#)). It might also be beneficial to include positively framed UCVs in local XAI tools because they are self contained within each feature explanation. Compared to framed visualizations, framed UCVs also require less effort to read. Most local XAI tools already use continuous prediction probabilities so implementing UCVs should be possible for different decision classes. To conclude, we feel that framing explanations with respect to the positive decision class will improve the understandability of local XAI interfaces and thereby, encourage users to interact with these interfaces.

## 7.2 Covariations in user belief, application domain and learning effects

When participants disagreed with predictions provided by the system, we found that understandability of our explanations changed. This aligns with the XAI study by [Springer and Whittaker \(2019\)](#) who found a similar observation when one participant in their study thought that the keyword "isolation" should have a greater contribution towards classifying a piece of text as negative (and thereby decreasing trust). However, this study was merely an observation made by a single participant. Nevertheless since we found that user belief does influence the understandability of our explanations, there might be confirmation bias ([Kahneman & Tversky, 1972](#)). In particular we found that when users disagreed with predictions, for positively framed visualizations, the positive UCV condition became more understandable while the negative UCV conditions became less understandable. Negatively framed visualizations remained somewhat similar. What this implies, is that when user belief is misaligned with a prediction, it is even more important to use positive UCVs compared to negative UCVs. In the end, confirmation bias is unavoidable since users will never agree with every single prediction. Our exploratory findings state that in these cases, it is even more important to use positively framed UCVs for local XAI tools. If this is not possible, it might be good to use positively framed visualizations. When users agree with the predictions, positively framed visualizations are still more beneficial than negatively framed visualizations.

We also found that our music AI system explanations were less understandable than our loan decision support system. More precisely, we found that positively framed visualizations were more understandable for the loan AI compared to the music AI. These results are in line with our argumentation of problem incompleteness ([Doshi-Velez & Kim, 2017](#)). We argued that a music system is trying to address a fundamentally more incomplete problem compared to a loan system. Although the Spotify features provide a reasonable explanation of the underlying song properties, there are still environmental, social and temporal variations that shape an individuals music preferences. This provides a level of *unquantified bias* that can't be captured by a system and might not be reflected in the explanations. Comparatively, our loan AI is less subjective and has a lower degree of unquantified bias. Banks usually have sufficient financial and personal information to reasonably capture if an applicant is eligible for a loan which is also reflected in loan explanations. Based on this argumentation, positively framed explanations in our loan decision support system might be more understandable than our music AI system.

However, in the first trial for each system, we found that there were no differences in understandability between the two application domains. Therefore upon first impressions, explanations might generalize across both domains. In line with XAI research, we argue that explanations are more relevant in contexts where users belief might not align with a given recommendation ([Miller, 2019](#)). In other words, explanations are more relevant when they provide answers to a contextual why question such as "why did the system give me this prediction that does not make any sense?". Therefore, in a more realistic application, explanations might only be presented in contexts where the user dislikes a prediction. If they need to be used more often, it should be done for domains that have a lower level of incompleteness (e.g. loan recommenders).

## 7.3 Limitations

The current study contains some methodological limitations. Firstly our duration measure is noisy which makes it less precise. This is because participants are presented with a lot of information and interaction options. In order to provide comprehensive explanations, we felt that this level of detail was



required. Nevertheless, noise could be introduced when participants clicked on description boxes, when they previewed songs in the music AI or took time interpreting feature values. In contrast, comprehension measures used in psycholinguistics generally manipulate more specific stimuli such as sentences where reaction times range between 750 to 1500ms (Kaup et al., 2006).

In the current study we ask participants understandability questions, a four item questionnaire and, a valence determination task. The combination of these tasks inherently puts cognitive load on participants. In the pilot study, participants stated that the repeated measures along with the valence determination task was quite tiring. This is because they are asked to mentally restructure feature explanations that contain multiple negations due to our framing manipulations. In our Literature review this does align with studies that found increased brain activity in participants processing negations (Carpenter et al., 1999; Herbert & Kübler, 2011). We also counterbalanced each condition so that participants didn't get used to a pattern (which didn't make the task any easier). In a more practical implementation, participants should be presented with an *option* to view the explanation. In the current experiment, participants were forced to review explanations even if they agreed with a particular prediction. Based on these observations, we think that the controlled environment might reduce the ecological validity of our findings. To apply our findings, future studies could test our framing manipulations in an ecologically valid system, such as a music recommender system.

Our study also had some technical limitations that might reduce ecological validity. For example, we don't use user profiles to make predictions. For the music AI we used a system that categorized songs from each genre based on Spotify song features. The explanations however, don't really provide participants with a clear link between their preferred genre and song features. In fact, there is no simple relationship because the decision tree classifier comes up with categorization values for each genre based on multiple decisions it makes for each separate feature. One alternative we considered was to show the typical feature values for each genre next to the explanation. This however would not allow us to fairly compare the music AI with the loan AI. Another technical limitation is that a typical music recommender system does not provide participants with songs that they dislike. Instead, it uses a prediction weights to provide recommendations that align with a participants user profile. Therefore, we can't be entirely sure that our findings generalize to recommender systems. However, for the purpose of comparing decision class with framing such similar to LIME visualization, we also needed to provide negative predictions. In the loan AI, the environment was artificial because we create a scenario where the participant assumes the role of a loan auditor. Although this allowed us to test a loan expert perspective, we think that the explanations would be more meaningful if the system provided predictions for each participant. This might reflect a real life financial situation where the need for explanations might be greater.

## 7.4 Implications and future research

As we have stated, our findings can help improve local XAI tools. To directly aid others in creating positively framed explanations, we created the ArgueView package. Currently, the ArgueView can be used in conjunction with LIME and SHAP. Qualitative user testing with positively framed explanations could add more depth to our study.

Furthermore, there is plenty of opportunity for future research. For example, the same study can be replicated with a more objective understandability measure like a test score (eg multiple valence determination tasks). However as noted in our pilot, such tasks are resource intensive and care must be taken not to overburden participants. To generalize our own findings, studies could look at how framed explanations influence system trust, transparency, recommendation acceptance and user satisfaction (Cramer et al., 2008; Knijnenburg, Willemsen, Gantner, Soncu, & Newell, 2012). Alternatively, future studies could investigate how visualizations using colored bars that symbolize positive (green) and negative (red) contributions influence understandability. Seeing as "wording-framed" visualization with neutral colors already help comprehension, framed colors could improve the comprehensibility even further. Furthermore, prototypes dedicated towards a particular domains could investigate how framing effects influence understandability in more specific contexts. For example, in a music recommender system, recommendations based on a user profile might provide more accurate suggestions that would improve user belief. In fact, a music recommender system based on user profiles might make our findings generalize across domains since the music problem becomes slightly more complete from a user's perspective (Doshi-Velez & Kim, 2017). In a loan recommender system, a user could fill in their own financial details so that the loan predictions become more personalized and the explanations become more meaningful. To minimize differences in understandability across domains, it would be beneficial to investigate when users require such explanations. For example, a study could look at the clicking and eye

tracking behavior of users interacting with a explanations after they agree or disagree with a prediction. In such a study one could also manipulate the explanation framing across different contexts (e.g. when users are confused or just curious about a prediction).



## 8 Conclusion

Despite the limitations, our findings provide conclusive answers to our initially posed research questions which we summarize in this section.

Our first research question aimed to examine how the addition of an underlying continuous variable influences the comprehensibility of an explanation. Our second research questions aimed to examine how a framed visualization and its alignment with the final prediction influences understandability. To answer the first question, we found that explanations containing positively framed UCVs were easiest to understand. To answer the second question, positively framed visualization were the easiest to understand. It is even more beneficial when positively framed UCVs are combined with positively framed visualizations. Although we find some evidence that explanation with visualizations framed with respect to the decision classes take lesser time to understand (eg LIME), we think that positive explanations are comparatively better. This is because local XAI tools might also contain negatively framed visualizations which we found decreases understandability. In particular, the current study found the explanations containing negatively framed UCVs along with negative visualizations to be the least comprehensible. Regardless of explanation framing however, there is some confirmation bias when it comes to understanding explanations. When there user agrees with a prediction it is even more important to use positively framed UCVs. When a user disagrees with a prediction, positively framed visualizations (without UCVs) are still beneficial. However, these explanations must only be used when the context is relevant. Otherwise learning effects could influence the effectiveness of the explanation based on the completeness of the problem that the system is trying to answer.

Although we provide conclusive evidence for our research questions, in order to verify the effect size and ecological validity of our findings, the study must be replicated and tested across different application domains and contexts alongside users. This study provides a starting point for further research.

## 9 References

- Boucher, J., & Osgood, C. E. (1969). The pollyanna hypothesis. *Journal of Verbal Learning and Verbal Behavior*. doi: 10.1016/S0022-5371(69)80002-2
- Carpenter, P. A., & Just, M. A. (1975). Sentence comprehension: A psycholinguistic processing model of verification. *Psychological Review*. doi: 10.1037/h0076248
- Carpenter, P. A., Just, M. A., Keller, T. A., Eddy, W. F., & Thulborn, K. R. (1999). Time course of fMRI-activation in language and spatial networks during sentence comprehension. *NeuroImage*. doi: 10.1006/nimg.1999.0465
- Clark, H. H., & Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology*. doi: 10.1016/0010-0285(72)90019-9
- Corblin, F. (1996). Multiple negation processing in natural language. In *Theoria*. doi: 10.1111/j.1755-2567.1996.tb00503.x
- Cramer, H., Evers, V., Ramlal, S., Van Someren, M., Rutledge, L., Stash, N., ... Wielinga, B. (2008). The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction*. doi: 10.1007/s11257-008-9051-3
- Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning*.
- Ferreira, J. J., & Monteiro, M. S. (2020). What Are People Doing About XAI User Experience? A Survey on AI Explainability Research and Practice. In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)*. doi: 10.1007/978-3-030-49760-6\_4
- Fischler, I., Bloom, P. A., Childers, D. G., Roucos, S. E., & Perry, N. W. (1983). Brain Potentials Related to Stages of Sentence Verification. *Psychophysiology*. doi: 10.1111/j.1469-8986.1983.tb00920.x
- Gilbert, D. T., Krull, D. S., & Malone, P. S. (1990). Unbelieving the Unbelievable: Some Problems in the Rejection of False Information. *Journal of Personality and Social Psychology*. doi: 10.1037/0022-3514.59.4.601
- Herbert, C., & Kübler, A. (2011). Dogs cannot bark: Event-related brain responses to true and false negated statements as indicators of higher-order conscious processing. *PLoS ONE*. doi: 10.1371/journal.pone.0025574
- Itkes, O., & Mashal, N. (2016). Processing negative valence of word pairs that include a positive word. *Cognition and Emotion*. doi: 10.1080/02699931.2015.1039934
- Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*. doi: 10.1016/0010-0285(72)90016-3
- Kaup, B., Lüdtke, J., & Zwaan, R. A. (2006). Processing negated sentences with contradictory predicates: Is a door that is not open mentally closed? *Journal of Pragmatics*. doi: 10.1016/j.pragma.2005.09.012
- Kess, J. F. (1992). *Psycholinguistics: psychology, linguistics, and the study of natural language*.
- Knijnenburg, B. P., Willemsen, M. C., Gantner, Z., Soncu, H., & Newell, C. (2012). Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, 22(4-5), 441–504. doi: 10.1007/s11257-011-9118-4
- Kuhlmann, M., Hofmann, M. J., Briesemeister, B. B., & Jacobs, A. M. (2016). Mixing positive and negative valence: Affective-semantic integration of bivalent words. *Scientific Reports*. doi: 10.1038/srep30718
- Kutas, M., & Federmeier, K. D. (2000). *Electrophysiology reveals semantic memory use in language comprehension*. doi: 10.1016/S1364-6613(00)01560-6
- Lai, V. T., Hagoort, P., & Casasanto, D. (2012). Affective primacy vs. cognitive primacy: Dissolving the debate. *Frontiers in Psychology*. doi: 10.3389/fpsyg.2012.00243
- Lazarus, R. S. (1984). On the primacy of cognition. *American Psychologist*. doi: 10.1037/0003-066X.39.2.124
- Lenzner, T., Kaczmirek, L., & Lenzner, A. (2010). Cognitive burden of survey questions and response times: A psycholinguistic experiment. *Applied Cognitive Psychology*. doi: 10.1002/acp.1602
- Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in neural information processing systems* (Vol. 2017-December).
- Matlin, M. W. (2016). Pollyanna principle. In *Cognitive illusions: Intriguing phenomena in judgement, thinking and memory*. doi: 10.4324/9781315696935
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1–38. doi: 10.1016/j.artint.2018.07.007
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should i trust you?" Explaining the predictions of

- any classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13-17-Aug*, 1135–1144. doi: 10.1145/2939672.2939778
- Sherman, M. A. (1976). Adjectival negation and the comprehension of multiply negated sentences. *Journal of Verbal Learning and Verbal Behavior*. doi: 10.1016/0022-5371(76)90015-3
- Silvera, D. H., Krull, D. S., & Sessler, M. A. (2002). Typhoid Pollyanna: The effect of category valence on retrieval order of positive and negative category members. *European Journal of Cognitive Psychology*. doi: 10.1080/09541440143000041
- Springer, A., & Whittaker, S. (2019). Progressive disclosure empirically motivated approaches to designing effective transparency. *International Conference on Intelligent User Interfaces, Proceedings IUI, Part F1476*, 107–120. doi: 10.1145/3301275.3302322
- Stumpf, S., Rajaram, V., Li, L., Burnett, M., Dietterich, T., Sullivan, E., ... Herlocker, J. (2007). Toward harnessing user feedback for machine learning. *International Conference on Intelligent User Interfaces, Proceedings IUI*, 82–91. doi: 10.1145/1216295.1216316
- Tintarev, N., & Masthoff, J. (2015). Explaining recommendations: Design and evaluation. In *Recommender systems handbook, second edition*. doi: 10.1007/978-1-4899-7637-6\_10
- Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist*. doi: 10.1037/0003-066X.35.2.151

## 10 Acknowledgements

This thesis has been a long yet rewarding process and many people have helped me along the way. This section gives me the opportunity to thank them.

Firstly, I would like to thank my supervisors Martijn and Sophia. Both of them have not only been very patient and understanding with me during difficult times, but also provided me with very valuable feedback that I will look back on in my future endeavours. Hopefully this study was helpful and I wish you both the best of luck in future studies.

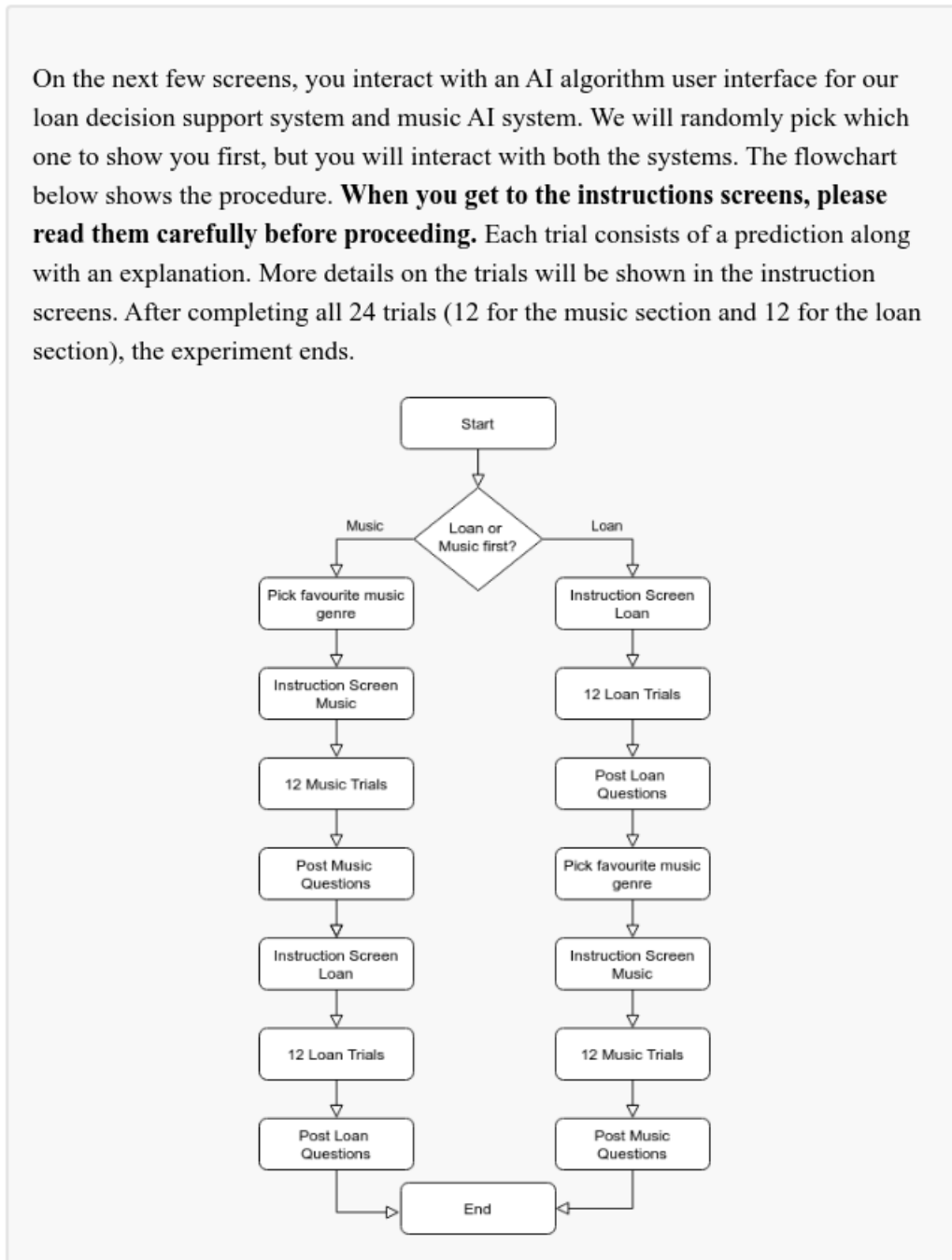
Secondly I want to thank all my friends who have been there for me in times of need during this pandemic but also during times of happiness.

Finally, I want to thank my parents and my counsellor who have unconditionally listened to, supported, and loved me throughout my entire journey.

## A Participant Instruction Screens

### Introduction

On the next few screens, you interact with an AI algorithm user interface for our loan decision support system and music AI system. We will randomly pick which one to show you first, but you will interact with both the systems. The flowchart below shows the procedure. **When you get to the instructions screens, please read them carefully before proceeding.** Each trial consists of a prediction along with an explanation. More details on the trials will be shown in the instruction screens. After completing all 24 trials (12 for the music section and 12 for the loan section), the experiment ends.



Continue

Figure 19: General introduction to our study

**Instruction Screen**

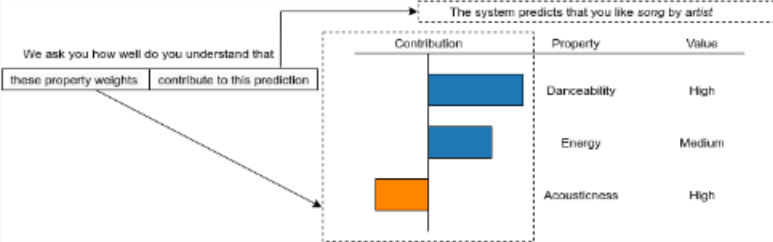
**The AI**

The AI system predicts whether you like the music based on a number of music characteristics such as the tempo, acousticness, energy and more. An explanation will be shown that displays why the AI system predicts you **like** or **dislike** a song. The explanation will indicate how these properties contribute towards the prediction. You will also receive song previews for each prediction.

**Setup**

You will first be shown six predictions along with their explanations. These predictions are made according to the genre you chose in the previous screen. You will then be asked to answer some additional questions. The explanations will then change and you will follow this procedure once more with different predictions. You will be informed when this happens. The music section ends after you complete all twelve trials and answer all questions.

**Task**



The visualization shows a prediction: "The system predicts that you like song by artist". Below this, a table lists properties and their values. To the left, a bar chart shows the contribution of each property to the prediction. The contribution is positive (blue) for Danceability and Energy, and negative (orange) for Acousticness.

Contribution	Property	Value
Positive (Blue)	Danceability	High
Positive (Blue)	Energy	Medium
Negative (Orange)	Acousticness	High

We ask you how well do you understand that these property weights contribute to this prediction.

For each trial it is your task to try to understand the explanation of the prediction, which is displayed as a set of labeled colored bars in the contribution column (see above). The blue and orange colors indicates how the value (eg. High) of a particular property (eg. Danceability) contributes towards the prediction (that you like or dislike the song). You have to indicate how well you understand this explanation. **Please respond to this question as soon as you have an answer. It does not matter whether you agree with the prediction or the values for each property.** On the next screen, we will ask you whether you agree with the prediction.

Continue

Figure 20: Music AI system introduction

**Instruction Screen**

In this section of the experiment, you will interact with our loan decision support system. It is your task to judge how understandable the presented explanations are.

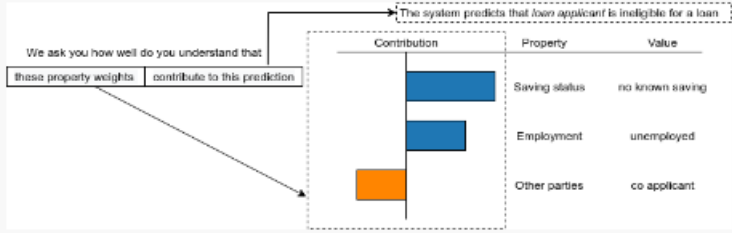
**Scenario**

In the upcoming screens you are asked to imagine that you are a loan auditor. A loan auditor is someone who reviews and verifies the accuracy of loan application decisions in a bank. The bank wants to implement a new loan decision support system. For each customer who applies for a loan, the system will predict where they are loan **eligible/ineligible** along with an explanation. You have to check that the explanations provided by the decision support system are understandable.

**Setup**

You will first be shown six predictions along with their explanations. You will then be asked to answer some additional questions. The explanations will then change and you will follow this procedure once more with different predictions. You will be informed when this happens. The loan section ends after you complete all twelve trials.

**Task**



The system predicts that loan applicant is ineligible for a loan

Contribution	Property	Value
[Large blue bar]	Saving status	no known saving
[Medium blue bar]	Employment	unemployed
[Small orange bar]	Other parties	co applicant

We ask you how well do you understand that these property weights contribute to this prediction?

For each trial it is your task to try to understand the explanation of the prediction, which is displayed as a set of labeled colored bars in the contribution column (see above). The blue and orange colors indicates how the value (eg. no known saving) of a particular property (eg. Saving status) contributes towards the prediction (that the applicant is eligible/ineligible for a loan). You have to indicate how well you understand this explanation. **Please respond to this question as soon as you have an answer. It does not matter whether you agree with the prediction or the values for each property.** On the next screen, we will ask you whether you agree with the prediction.

Please read each explanation carefully before responding. Remember there is no right or wrong answer. Good luck!

Continue

Figure 21: Loan decision support system introduction

## B Plots for Testing Regression Assumptions

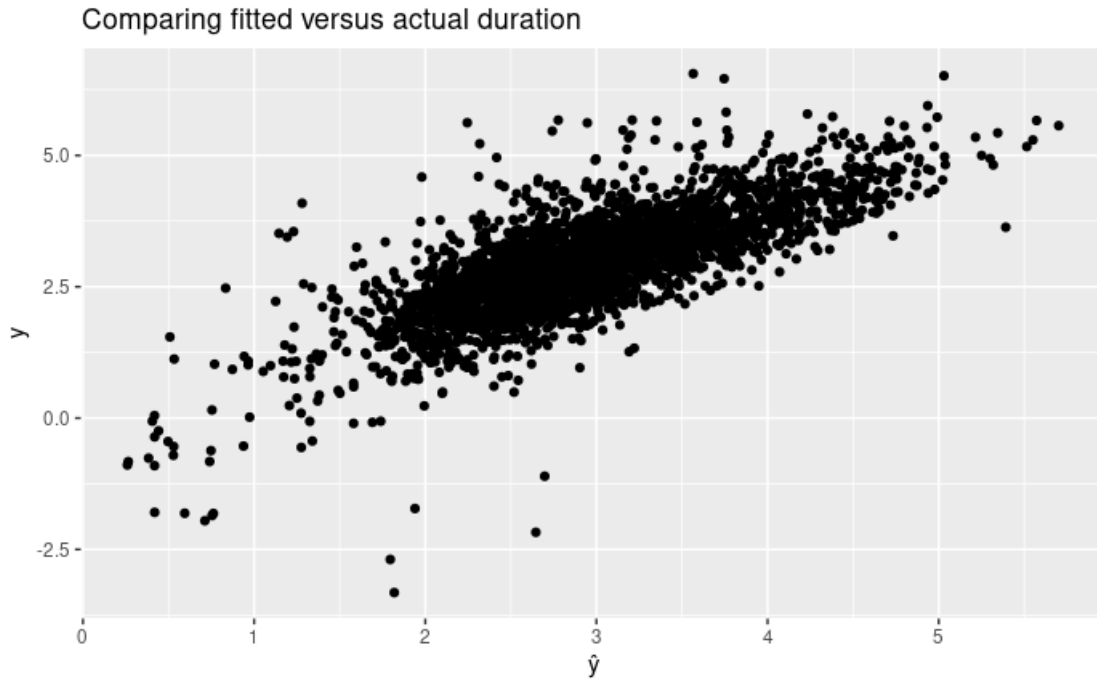


Figure 22: Duration model outliers comparing fitted prediction with actual measures. For the five outliers, the fitted actual measures are between -1.5 and -2.5 while the model predictions are between 1.5 and 3.

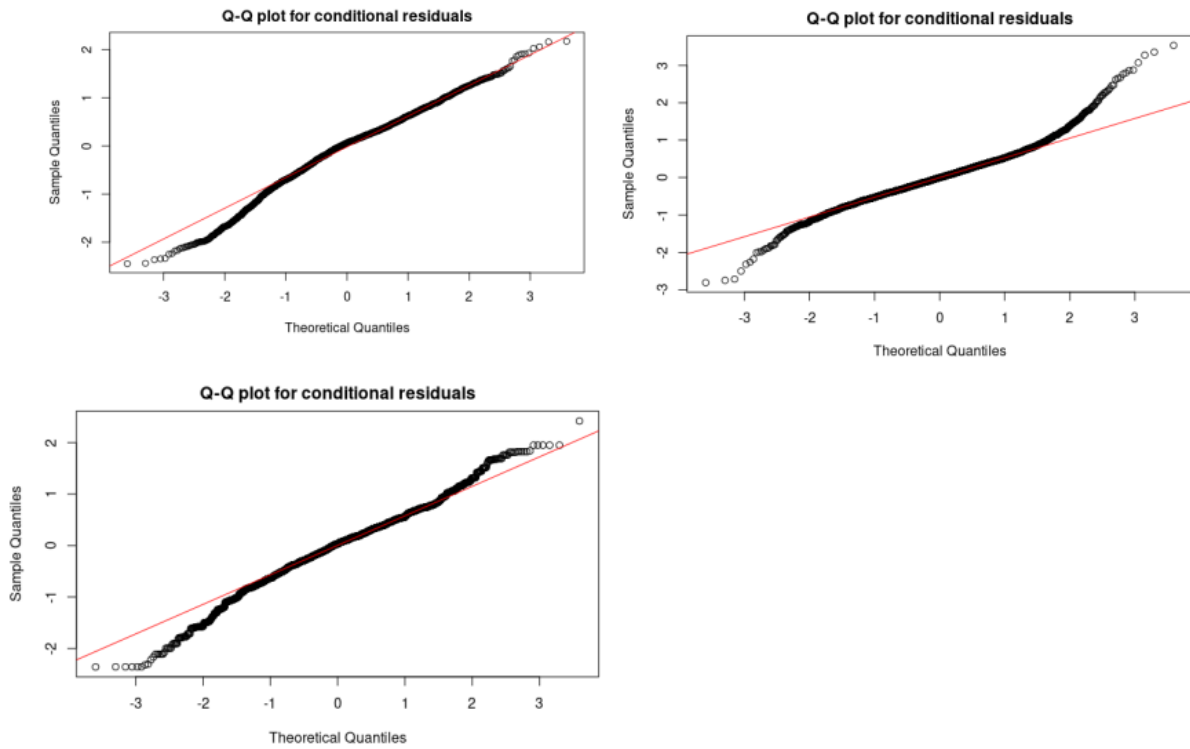


Figure 23: Q-Q plot of conditional residuals for Understandability (top left), Duration (top right) and Perceived Understandability (bottom left). For most data points, the residuals fit the linear curve.