UNIVERSITY OF TARTU
Faculty of Science and Technology
Institute of Computer Science
Computer Science Curriculum

Rasmus Moorits Veski

Measuring Human Preferences in Counterfactual Explanations

Bachelor's Thesis (9 ECTS)

Supervisors: Marharyta Domnich, MSc
Kadi Tulver, PhD
Raul Vicente, PhD

Tartu 2024

Measuring Human Preferences in Counterfactual Explanations

Abstract. With modern machine learning models' decision-making process growing increasingly complex, the reasoning behind their decisions becomes more opaque. An effective method to understand why a model made a specific choice is through counterfactual explanations. However, this raises another challenge: how to produce explanations that are most useful for humans. One possible approach is to incrementally integrate human cognitive biases into counterfactual search algorithms. To investigate which biases are relevant, this thesis conducts a survey in which respondents rate counterfactual explanations based on overall satisfaction and adherence to seven explanatory criteria. The measured biases provided insights into the role of sub-criteria in assessing the subjective measure of overall satisfaction. However, data analysis on the responses indicated, that the biases are strongly interwoven, with fewer underlying factors possibly accounting for human biases. Overall, humans seemed to place the most emphasis on the feasibility of the explanation. The findings in this thesis and the dataset generated from the questionnaire help pave the way towards developing more human-like explainable systems.

Keywords: Explainable AI, Counterfactual Explanations.

CERCS:P176 Artificial intelligence

Inimlike eelistuste mõõtmine kontrafaktuaalsetes selgitustes

Lühikokkuvõte: Masinõppemudelid muutuvad kiire arengu ja uute struktuuride lisamise tõttu üha keerukamaks, mistõttu on nende tehtud otsuste taga olevad protsessid kasutajale vähem läbipaistvad. Üks võimalik viis mõista mudeli tehtud otsuseid on luua nende kohta kontrafaktuaalseid selgitusi(*counterfactual explanations*) kontrafaktuaalsete selgituste genereerimisalgoritmiga. See meetod aga tõstatab omaette probleemi: kuidas moodustada kontrafaktuaalseid selgitusi, mis on inimestele kõige kasulikumad. Üks variant on kontrafaktuaalsete selgituste genereerimisalgoritmidesse inimlikud eelistused sisse kirjutada. Uurimaks mida inimesed selgituste puhul hindavad, viiakse selle töö raames läbi küsimustik, kus osalejad hindavad kontrafaktuaalsete selgituste üldist kvaliteeti ning vastavust teatud seletatavuse väärtustele (*explanatory virtues*). Küsimustiku vastustel rakendatud andmeanalüüs osutas võimalusele, et mõõdetud seletatavuse väärtused on omavahel tihedalt seotud ning potentsiaalselt saab mõõdetud väärtused koondada väiksemale arvule faktoritele. Üldiselt väärtustasid inimesed enim selgituste teostatavust (*Feasibility*). Küsimustiku raames loodud andmestikku ning andmeanalüüsi avastusi saab kasutada tulevaste kontrafaktuaalsete selgituste genereerimisalgoritmide inimsõbralikumaks muutmisel.

Võtmesõnad:
Selgitatav tehisintellekt, kontrafatuaalsed selgitused

CERCS:P176 Tehisintellekt

Contents

1 Introduction

With artificial intelligence (AI) and machine learning (ML) playing an increasing role in everyday life, it is vital to be able to trust its decisions. If a self-driving car decides to turn off the road and run into a person, the root of the decision needs to be investigated to see if the decision-making process can be corrected. With the continuous upscaling of ML models and the introduction of increasingly intricate model architectures such as transformers, the final models users interact with are becoming so large and complex that even experts cannot understand all decision-making pathways the model can take.

A possible solution to understanding AI-made decisions is counterfactual explanations: the model providing an example of a similar input to the one given, which would result in a different output from the original. There are algorithms made for generating such explanations, such as DiCE [1] or FACE [2]. Easily readable explanations describing why the model made a certain decision are already a major step towards understandable AI, but they could be further enhanced.

There is evidence of humans having preferences towards certain underlying features in counterfactual explanations [3]. Therefore, counterfactual explanations produced by machine learning models, which are used by humans, should align with these discovered human preferences. To achieve this, counterfactual explanation algorithms need to be designed with human preferences in mind.

However, there is no consensus about what kind of explanations humans prefer. Furthermore, there is no dataset with humans evaluating explanations based on explanatory virtues to draw these conclusions from. Without definitively knowing what humans desire, integrating human wants into counterfactual algorithms would be premature. To generate a dataset containing explanations and their inherent values, people's thoughts about explanations need to be analysed. This can be achieved through a questionnaire where people can evaluate explanations on overall satisfaction and adherence to criteria (metrics) representing human-held values.

Through statistical and machine-learning methods applied on the responses of the questionnaire the main goal of the thesis is to find evidence about what explanation qualities are most relevant for humans. To this end, the following subquestions are researched:

- How many metrics are sufficient for predicting human satisfaction?
- Were the chosen metrics representative of human preferences in explanations?
- Can satisfaction be predicted from the chosen subset of metrics?

In Chapter 2, this thesis will outline the previous work done surrounding explainable AI models and human preferences in counterfactual explanations. Chapter 3 will give an overview of a survey conducted to determine what humans value in counterfactual explanations. Chapter 4 gives a brief rundown of methods used for data analysis. Chapter 5 will analyse the results of the survey in terms of the interconnection of metrics and their significance towards overall satisfaction. Finally, Chapter 6 will discuss the impact of the results of the questionnaire on current and future counterfactual algorithms.

Appendix I consists of all the questions in the created questionnaire. Appendix II links to the code used to analyse the results of the survey. Appendix III. contains figures and tables illustrating the general statistics and data distribution of the survey's responses. Appendix IV. contains the training info and results of the machine-learning models applied to predict satisfaction from the measured metrics.

Artificial Intelligence tools ChatGPT 3.5 [4] and Microsoft Copilot [5] were used to enhance wording in some sections of the thesis.

2 Background

The use of machine learning and AI is skyrocketing and does not show signs of stopping [6]. Computer vision algorithms recognise faces to unlock devices, Large Language Models (LLMs) provide answers to queries, personalised algorithms recommend a variety of services, etc. The majority of users have no idea how the models providing these services work or how the models "think". Most people do not fully trust AI [7], however, the large number of users LLMs such as ChatGPT have [8] indicates that a significant number of people find AI useful enough to use.

These models should not be accepted just because they happen to work most of the time. For example, ChatGPT has produced prejudiced texts [9] and facial recognition makes mistakes under suboptimal circumstances [10]. These systems are not fail-safe. To have trust in models' decisions, there needs to be a deeper understanding of how they make them. To achieve this, multiple methods have been developed. An overview of the most relevant methods will be given in the following sections.

2.1 Explainable AI

The problem with understanding how modern machine-learning models operate lies in their complex construction. Many models function as a black-box, meaning the user cannot see the mechanics behind deducing an output from the input. For many end-users, all they see when utilising a model is entering some input and getting a number out, as depicted on Figure 1. The output may make sense, but the process of reaching it and the reasons are hidden.

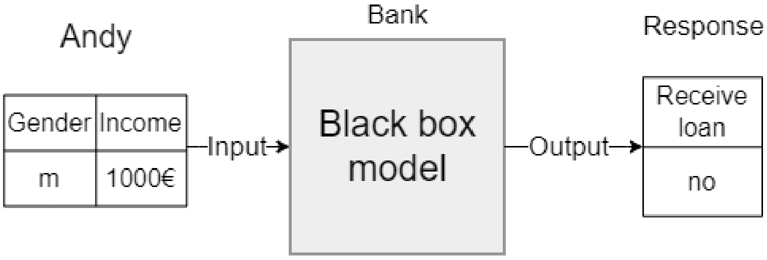


Figure 1. Layout of a black-box model producing an output from an input. In this example a person Andy applies for a loan with his data and the model returns a negative output.

Receiving responses without any reasoning as Andy did in Figure 1 can be frustrating and lower trust in the model itself. If there is no way to explain why a model made a decision, that decision might not be accepted or trusted.

To provide insight into these black-boxes, the field of Explainable AI (XAI) was created. There are multiple techniques for making sense of how a model makes decisions. One approach is to use tools that make the whole model more interpretable. For example, LIME (Local Interpretable Model-Agnostic Explanations) [11] can be used to build up knowledge about a model's decision boundary with permutations to a single prediction, resulting in an understandable proxy model explaining the more complex model, which LIME was used on. SHAP (SHapley Additive exPlanations) [12] can be used to deduce how important features are for final predictions. Furthermore, when utilising feature info from multiple explanations it is possible to gain general understanding of what the model takes into account when making decisions. Models can also be built to be inherently understandable, such as the Decision Tree, where the process of making a prediction can even be drawn out [13]. All of these methods provide a better overview of how a model makes decisions. This thesis will focus on a different, instance-based approach: counterfactual explanations.

2.2 Counterfactual Explanations

Given a trained model and a classification task, the model takes an input and gives its prediction for the output. When these predictions are used automatically and churned out thousands every second, as is the case in some tasks, such as a self-driving car detecting the location of the edges of the road, then the details of the decisions tend to get overlooked. However, the moment one of these decisions affects someone's livelihood or well-being, that person might want to know what prompted that decision.

A classic example is a model deciding if a person receives a loan from the bank. Suppose there is a person named Andy, who is a 22-year-old man, who makes 1 000€ per month. Andy applies for a loan and the bank's AI rejects him. Andy would probably like to know why such a decision was made. The bank could reply that if Andy made 1500€ per month, he would get the loan he wished for.

Table 1. Example of factual and counterfactual in a tabular setting

	Name	Gender	Income	Received_loan
FACTUAL	Andy	m	1000	no
COUNTERFACTUAL	Andy	m	1500	yes

That was an example of a Counterfactual Explanation (CE) or counterfactual. A CE is an explanation with the structure "if X had occurred instead of Y, then Z would have happened instead of W". In the context of machine learning, a CE highlights necessary changes in the input values or factual for the model to give a different output. Table 1 is an example of how a switch in the income column can prompt a change in the Received_loan column, which means Andy would have got the loan.

2.2.1 Formalization of Counterfactual Explanations Objective Function

For the purpose of creating CEs from machine learning models, numerous counterfactual search algorithms (CSAs) have been created. A CSA's application is to take a trained model and an input vector and use a search algorithm to find a similar vector to the original input, which would have a different output if given as input to the model. Figure 2 depicts how Andy from the previous example (Table 1) might use a CSA to find what he needs to do in order to get a loan.

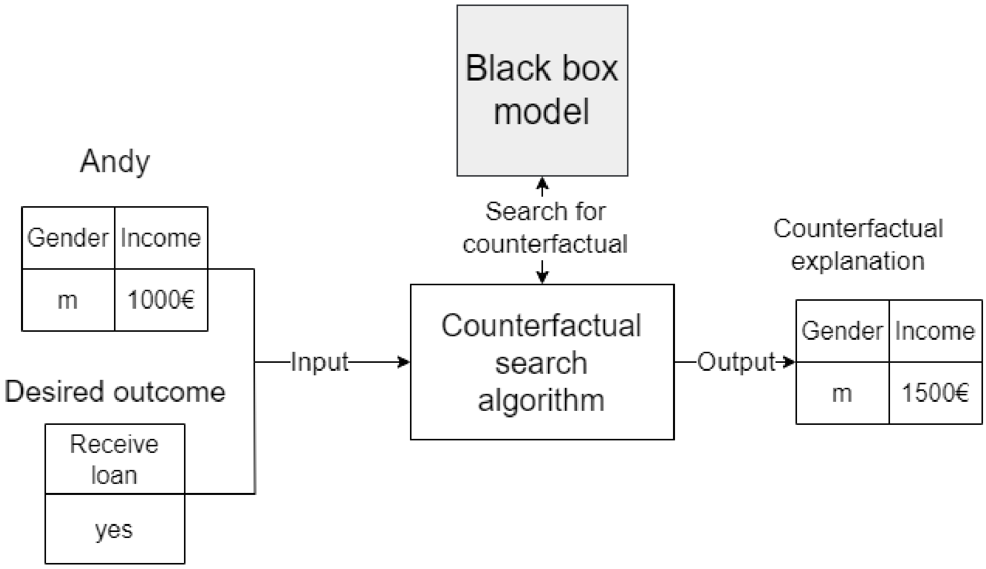


Figure 2. Graph of how Andy (Table 1) might use a counterfactual search algorithm to find what he needs to do in order to get a loan. He inputs his data just like in Figure 1 in addition to what he wants. The CSA finds a data point similar to the original which would have been granted a loan by the model.

There are multiple methods developed to search for counterfactual explanations. The process for identifying counterfactual point, as shown in Figure 2, may suggest modifications to different features or varying magnitudes of change to the same feature. The method by which an algorithm transitions from one decision space to another can vary significantly. An example of the decision boundary that the CSA has to cross can be seen in Figure 3.

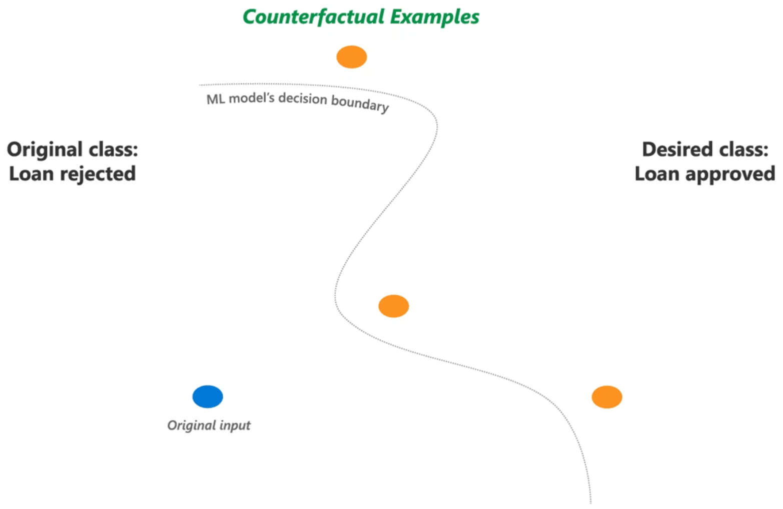


Figure 3. Example of a decision boundary between a factual scenario and a counterfactual scenario as portrayed by Mothilal et al. [14].

As seen in Figure 3, there is a (possibly) endless amount of counterfactuals on the other side of the decision boundary. How to choose the best counterfactual explanation for the given factual also needs to be decided by the search function. To formalize the search for a counterfactual in the case of binary classification, let's denote f as a trained black-box ML model that maps the input space $\mathcal{X}$ to the output space $\mathcal{Y}$. In a scenario where original point x generates prediction $f(x) = 1 - y$, we search for counterfactual point x', such that $f(x') = y$ [15]. Thus, to obtain counterfactual point, the following function should be optimized:

$$\arg\min_{x'} loss(f(x'), y) + dist(x', x), \tag{1}$$

where:

- loss$(f(x'), y)$ - loss function that validates if the counterfactual outcome is equal to the desired outcome.
- dist(x', x) - distance function between the original point x and the counterfactual point x'.

Equation 2.2.1 is an example of a simple search function, loosely based on the search function by Wachter et al. [16]. The goal of a search function is to find a counterfactual x', that has the minimal distance to the factual x and the minimal loss between the target and the model's prediction of the counterfactual. Equasion 2.2.1 serves as a basis, where specifications can be added. Both the loss and distance function can be customised and additional clauses can be added.

2.2.2 Selecting Distance Function for Counterfactual Optimization

The search function determines the nature of the counterfactual that is generated from the factual data. However, how to make the best counterfactual from a factual is not a straightforward problem. Karimi et al. have even suggested that finding the best counterfactual is an NP-complete problem [17]. Firstly, it needs to be determined what are the desired properties for a counterfactual. The most straightforward method to assess the quality of a counterfactual is to measure the counterfactuals' distance from the factual in the feature space.

There are distinct distance function for two different data types: continuous and categorical. **Continuous** data is data, that has an infinite number of possible values in a given range. The most commonly used distance functions for this type of data are Euclidean L_2 and Manhattan L_1. Weighted Manhattan distance was also recommended by one of the first papers discussing counterfactual algorithms by Wachter et al. [16]. They suggested it to be a suitable starting measurement, which can be further improved with specific domain knowledge. This rating could be effectively utilised in a an automated industrial setting, minimising material cost, distance, etc.

Categorical data is data that can't be numbered or counted. It further splits into nominal data, which can be ordered (e.g. education) and nominal data, which can not be ordered (e.g. colour). Distance functions for this data type are inherently more difficult, as the data values themselves cannot always be compared in magnitude. To compare the difficulty Domnich and Vicente [15] in their paper utilise the weighted L_0 norm, which checks the equality of categories values with identity function, which is normalized with number of categories.

While distance is important and should minimal, it is not sufficient for most real-world applications. A simple example to illustrate the shortcomings of distance is a situation, where Andy from the previous example (Table 1) is told he has to be 20 years older in order to get a loan. Without assigning any weights, the distance from the original factual is less than asking Andy to earn 500€ more per month. In reality, it is more difficult to actively change your age than your income.

An alternative to distance that better mirrors the real world is cost. Cost can tax changing (certain) attributes in certain ways more harshly. Ustun et al. propose multiple possible cost functions in their paper. They suggest utilising a logarithm function for cost, so as to make changing values into higher and lower percentiles of the distribution more costly than inside the middling percentiles. The aim is to account for the fact that it is rarer for a value to be in outer percentiles and therefore more difficult to reach [18]. This approach directs the algorithm to prefer making more changes towards common values. A switch from middle school education to high school education would be less costly than from a Master's degree to a PhD.

In addition to the search function, many CSAs have extra tools to further customise the counterfactual produced. The algorithm can be told to change features only within the known value domain, within set boundaries or refrain from changing certain features at all. While these extra tools can be used to successfully generate human-preferred counterfactual explanations with known datasets and someone setting the rules, this thesis will focus on the possibility of having human-preferred generalisations for the entire algorithm with no manual parameter tuning required.

The discussed distance and cost functions[15, 16, 18]may help produce better CEs in some aspects, but a definitive all-encompassing search function accounting for all human needs is likely as complex as humans themselves. This thesis will discuss potential directions and ideas for tuning cost functions towards what humans prefer.

2.3 Human Preferences in Counterfactual Explanations

In the context of this thesis, the goal is to mould the CEs into something that humans find useful. This requires understanding what humans wish to see in CEs, which is a problem that cannot be definitively solved due to the diverse views and preferences of humans, but can be approximated through studies in psychology.

Ruth M.J. Byrne lists multiple ways how people think counterfactually in her paper [3]. The most relevant to this thesis is her finding that humans tend to consistently mutate factuals to counterfactuals in particular ways. Specifically, she states that humans, when imagining counterfactual scenarios, change features and feature states that:

- are exceptional within the spread of data;
- are under the control of the protagonist (themselves);
- have happened recently;
- can be influenced through action rather than inaction.

Other relevant findings listed by Byrne are that humans tend to generate additive counterfactuals, trying to go beyond and add to the information given, rather than tinkering with the known information. Another observation Byrne brings up is that humans construct counterfactuals more easily, if the original factual has a negative outcome [3]. This might be encoded in humans for practical reasons - why try to make an already good situation fail?

Tim Miller lists four major findings for general explanations in his article [19]. All four point to the same phenomenon: explanations are contextual. No explanation can be considered "good" in a vacuum. While there are purely mechanical properties of explanations such as length that can be measured, a person only considers an explanation good if they believe that it addresses the problem and that it falls in line with their (or their peers') beliefs.

An important difference between counterfactual algorithms and humans is that humans can use the entirety of their life experiences to make decisions. Trained models make decisions solely based on given datasets and boundaries. Models cannot move outside of their domain. Fully bridging this gap without general artificial intelligence is likely impossible, but there are ways to try and bring these decision-making situations closer.

One is to attempt to divide human preferences about counterfactuals into subcategories. As Byrne and Miller both brought up, the usefulness of counterfactual explanations depends on multiple factors [3], [19]. Hoffman et al. propose two collections ("goodness checklist" and "explanation satisfaction scale") of questions to grade how well the explanations line up with certain human expectations [20]. The questionnaire used for this thesis will have a similar base list.

These preferences can then be implemented into algorithms one by one. There have been numerous proposed divisions of these sub-preferences. This thesis will explore the effect of some of the proposed preferences on how humans evaluate counterfactual explanations described in the following sections.

2.3.1 Feasibility

An explanation is feasible if the actions suggested by the explanation are practical, realistic to implement and actionable. Feasibility measures if the reader thinks the explanation is doable in reality. For example, in Andy's case (Table 1), asking him to earn 10000€ per month would probably be considered infeasible for most, as he would have to increase his salary tenfold.

Feasibility is one of the more studied proposed metrics. Karimi et al. [17] use a similar term for a concept falling under the definition used in this study: actionability. Moreover, in their paper they expand on the idea of actionability stating that non-actionable features can be both mutable (e.g. age) and immutable (e.g. birthday). In their table of existing counterfactual search algorithms, many integrate these biases into their decision-making process of the algorithm. In this thesis, these two subcategories will both be counted under Feasibility, as both sub-definitions indicate non-actionability and the goal is to create counterfactuals that people find useful. While some infeasible features can conceptually be mutated, they are equally immutable for someone who needs to follow the explanation.

2.3.2 Consistency

An explanation is consistent when parts of the explanation are logically coherent and do not contradict each other. If an AI recommends two changes that oppose each other, then the explanation cannot be fully acted on or the effect of the explanation is lessened.

If Andy (Table 1) was told to earn more money but at the same time to quit his job and change his work position to jobless, that could be considered contradictory.

How Consistency appears in explanations has not been fully agreed on in literature. In their article, Zemla et al. separate "internal coherency" and "external coherency" [21]. Internal coherency is whether parts of the explanation are coherent among themselves, while external coherency is the explanations' coherence with general known facts about the world. This thesis will solely focus on what Zemla et al. called internal coherency, which in this thesis is called Consistency. The reason for this is that external coherency falls under what this study counts as Feasibility (Section 2.3.1) and/or Trust (Section 2.3.4). When part of an explanation goes against what is known to be possible in the outside world, then achieving what the explanation instructs is not possible (infeasible). On the other hand, when the explanation instructs something possible, but clearly against the direction of achieving the outcome, users would have little trust in the explanation successfully explaining how to reach a different result.

Karimi et al. talk about Consistency in the context of how plausible the location of a counterfactual is in the data distribution, separating domain-consistency, density-consistency, and prototypical-consistency [17]. Their definition of Consistency has less to do with how humans sense consistency in explanations and more about the data spread around the counterfactual. While plausibility is a useful tool towards creating actionable counterfactuals through having counterfactuals in denser regions of data, the survey in this thesis does not ask participants to measure plausibility, as they do not have knowledge of the data distribution.

A more similar Consistency definition to this thesis is used by Rasouli and Yu in their paper about their CSA CARE [22], where they define coherency as retaining the correlation between features.

2.3.3 Completeness

An explanation is complete if it is sufficient in explaining the outcome. Completeness captures the reader's sense that something is missing from the explanation. If the bank asks Andy (Table 1) to raise his credit score, explanations offering concrete steps to do so would probably feel more complete.

Completeness in an explanation has often been discussed in the context of causal graphs. A causal graph is a graph in which the nodes are features and the edges are the relationships between the edges. An explanation would be perfectly complete if the entire journey along the graph from changing some features to the target feature is reflected in the explanation, and all the incoming edges into the target are utilised.

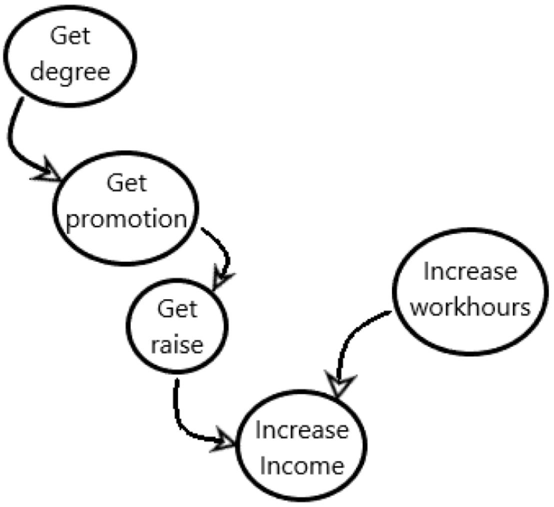


Figure 4. Example of a causal graph which can be used to determine the Completeness of a an explanation directing how to increase income.

Human thinking and causal graphs might not overlap fully. Some logical leaps are more obvious than others. Using Figure 4 as a reference, suggesting someone to earn more money without telling them to get a raise at their job is probably a more glaring hole in completeness than not specifying the extra qualifications they need to get. Similarly, the reader might or might not sense the necessity of increasing work hours, when a raise is already present in the explanation.

Zemla et al. found in their article [21], that at least in the case of everyday explanations, their participants detected when there were gaps in the explanations and rated explanations with gaps significantly lower. Furthermore, they found that the length of the completeness chain did not have a significant effect on the Satisfaction score, only whether that chain had holes in it.

2.3.4 Trust

A person has high Trust in an explanation, if they believe that the suggested changes would bring about the desired outcome. Through this metric people can express whether they believe that following the explanation would solve their problem. If the explanation offered to Andy (Table 1) would recommend earning less money to get a loan, it would be hard to believe that it solves the problem.

Out of all the metrics, Trust is probably the most connected to the machine learning model and data that the counterfactual algorithm uses. When the data is valid and the model is accurate, the counterfactual algorithms will inevitably construct scenarios in which the goal is reached.

In prior literature, Trust has sometimes appeared in a different context than the one in this thesis. Rather than asking respondents if they trust that the explanation leads to the desired outcome, it is asked if they trust the system that made the decision. This is a significant difference as from specific instances the respondents would have to decide if they trust the model as a whole. In their paper, Hoffman et al. even have a line of questions designed to measure the Trust that people have in XAI models [20]. While this is also a relevant direction of research to find ways to make people trust models more and to generate explanations that create a favourable outlook towards AI, this thesis's definition of Trust is independent of any model. This thesis evaluates explanations only on the basis of the explanation itself, focusing solely on improving their intrinsic value for humans.

2.3.5 Understandability

An explanation is understandable to a person if they fully understand the wording and have no doubts about its meaning. If the bank used confusing phrases or overly specific terminology that Andy (Table 1) does not know, he would consider the explanation to have poor Understandability.

Understandability is unlike the other mentioned values, as it is more a prerequisite to even utilise an explanation. What a person understands depends on their knowledge about the domain and mastery of the language the explanation is in. As Understandability is such an integral part of using any text, this metric serves more as a method to gauge whether the respondents in the questionnaire are even able to grade the other metrics.

2.3.6 Fairness

An explanation is fair if it is unbiased for different user groups and does not operate on sensitive features or induce bias. If the bank told Andy (Table 1) to change his gender to receive a loan, it would be highly discriminatory.

What a person considers fair is of course dependent on their value system. The notion of what is socially acceptable and what is not has been malleable through history and across cultures. Therefore algorithms incorporating Fairness is a sensitive subject, as there likely is not a global consensus about what is fair. With the amount of background knowledge required to decide what features are not acceptable to change, it seems unlikely that this sort of information could ever be encoded into an algorithm. Having the chance of a CSA making offensive suggestions may prompt considering leaving sensitive info out of training models altogether. However, models are not trained only to have counterfactuals created from them. Sensitive info is still valuable in some models and leaving it out may prove more harmful. For example in medicine, age, race and gender are all very relevant when testing for diseases. The incorporation of Fairness in counterfactual search algorithms must come from the algorithms themselves.

Fairness in machine learning is a particularly delicate matter. A common criticism is automated speech transcription working with different accuracies for different demographics, demonstrated by Koenecke et al. in their article [23]. Shams, Zowghi and Bano in their literature review brought out that while authors discuss diversity and inclusivity in their papers, actual solutions are rarely brought forward [24]. For ethics' sake, it is important to research how humans perceive Fairness in counterfactuals.

2.3.7 Complexity

An explanation has perfect Complexity if it has the appropriate level of detail and depth. In contrast to other metrics, Complexity has the optimal value in the middle. An explanation can be too simple, like telling Andy (Table 1) to "earn more money" instead of a concrete number. Explanations can also be too complex, like arranging Andy's whole life minute-by-minute in order to increase his salary. An explanation with ideal Complexity has an appropriate amount of feature changes and gives the user all the necessary information without any redundancy.

From its definition, Complexity possibly has the highest degree of variance from person to person. What explanation a person considers sufficient can derive from how much they care about the subject or how knowledgeable about the domain they are. A person who knows less about the goal they want to reach might wish for more details, while an expert wants as little hand-holding as possible, as they can deduce all else.

Complexity is tightly connected to Sparsity, which refers to the amount of changes in the explanation. Generally, the longer the explanation the more complex it is. Historically there has been evidence that simpler and shorter explanations are better [25, 26]. Many CSA studies research whether the algorithm accounts for Sparsity [17, 27]. In more recent years though, there have been results suggesting that this notion is false [28], [21]. Karimi et al. have suggested that rather than pursuing algorithms built towards generating sparse counterfactuals, it might be better to impose it as a constraint, limiting the counterfactuals to a certain number of feature changes [17].

2.4 Incorporating Human Preferences in Counterfactual Algorithms

There have already been strides towards incorporating certain human preferences in counterfactual search algorithms. Karimi et al. have gathered a comprehensive list of dozens of CSAs and assembled them into one table, categorising the algorithms by what biases are built into them (among other things). The inherent values in explanations they studied were Actionability (split into unconditional and conditional), Plausibility (split into domain-consistency, density-consistency and prototypical-consistency), Diversity and Sparsity. While the number of algorithms studied is vast and varied, there is no one algorithm able to incorporate all [17].

A similar table amassing multiple counterfactual search algorithms is provided by Verma et al. [27]. Rather than focusing on the inclusion of explanatory values like Karimi et al. [17], Verma et al. look directly at how the algorithms function, putting emphasis on how the algorithm interacts with the model and the data. A notable inclusion in their list of criterions is how the algorithms deal with categorical features, which is one of the more difficult steps towards human-preferred counterfactuals. Once again there are algorithms accounting for each criterion, but no one algorithm accounts for all [27].

Many CSAs have been advertising themselves as having solved certain areas of human preferences. An example of applying Feasibility can be found in Poyiadzi et al.'s algorithm FACE [2]. In their paper, they propose an algorithm which is aware of the underlying data distribution. Because of it, it can propose the change from a factual to counterfactuals that lie in high-density regions of the data. Generating counterfactuals that lie near other data points guarantees that the specific outcome is achievable. Furthermore, FACE's search function is based on graph traversal using Dijkstra's algorithm through dense regions of data. This setup aims to produce counterfactuals by making lots of little changes, hoping that every jump from data point to data point is feasible due to the proximity [2]. This algorithm is a great step towards suggesting changes that lie in normal boundaries but does not remedy the problem of having to scale categorical features to compare them with numbers. How big is a change in work position from cashier to store manager, when translated to change in income? Furthermore, it does not account for features' directions. Changing some features (e.g. age) in certain directions can also be considered infeasible.

Rasouli and Yu attempt to incorporate multiple human biases in their CSA CARE [22]. Specifically CARE proposes a modular approach to account for metrics such as Validity, Soundness, Coherency and Actionability. The validity module makes sure that the CE is the desired class, as close to the factual as possible and as long as possible (for interpretability purposes). Soundness checks that the CE lies close to other data instances with the same class as the CE and that the path from factual to CE is connected with a continuous path (similarly to FACE [2]). The way CARE incorporates Coherency is to first measure underlying correlations between metrics and then when presented with a factual, try to retain the correlation balance within the counterfactual. While this certainly eliminates two connected features moving in opposing directions and thus the chance of bad coherency, this method comes with lots of dataset-specific nuances and may present unwanted effects. Namely, when two features are not be directly correlated in real life but happen to be correlated with the dataset. Suppose there is a correlation between income and age in a dataset. When the model wants to recommend increasing income it also has to increase age to keep the correlation stable. While it may be realistic that with higher age come higher wages, increasing age is hardly actionable. To remedy this, in the actionability module users can apply domain-specific constraints to make the algorithm refrain from changing chosen features. While this works, the ideal is to have a model with no need for human effort.

An alternative to better search functions is to generate a maximally diverse set of different CEs for one factual and then let the user pick what explanation they like. This path has been researched by Wachter et al. [16], Mothial et al. [1], Sharma et al. [29] and others. Diversity is great when user preferences are not known or when there are multiple good alternatives. However, this thesis focuses on the possibility of having CSAs generate useful counterfactuals through algorithmic quality, not through having the most options.

3 Creation of Questionnaire

To deduce what clauses are worth integrating into counterfactual algorithms, what humans prefer needs to be extracted from humans themselves. To this end a questionnaire was created in collaboration with Julius Välja [30] and the University of Tartu Natural & Artificial Intelligence Lab (NAIL). The questionnaire contained 30 factual-counterfactual pairs, which were graded by participants in accordance with chosen metrics and overall Satisfaction. The goal is to utilise the responses to deduce what metrics are important for people in order to focus on integrating these preferences into counterfactual algorithms.

3.1 Metrics Measured in Questionnaire

The metrics used for the study were chosen through review of previous papers with a similar requirement of dividing human preferences. After thorough consideration, seven metrics that sufficiently covered most aforementioned human biases were selected, listed in Table 2.

Table 2. Evaluation criteria used in the questionnaire. Table adapted from Välja's thesis [30].

Metric	Definition	Rating scale
Overall satisfaction	This scenario effectively explains how to reach a different outcome.	6-point Likert scale, from 1 to 6
Feasibility	The actions suggested by the explanation are practical, realistic to implement and actionable.	6-point Likert scale, from 1 to 6
Consistency	All parts of the explanation are logically coherent and do not contradict each other.	6-point Likert scale, from 1 to 6
Completeness	The explanation is sufficient in explaining the outcome.	6-point Likert scale, from 1 to 6
Trust	I believe that the suggested changes would bring about the desired outcome.	6-point Likert scale, from 1 to 6
Understandability	I feel like I understood the phrasing of the explanation well.	6-point Likert scale, from 1 to 6

Fairness	The explanation is unbiased towards different user groups and does not operate on sensitive features.	6-point Likert scale, from 1 to 6
Complexity	The explanation has an appropriate level of detail and complexity - not too simple, yet not overly complex.	5-point Likert scale, from -2 to 2

Table 2 summarises the names of the used metrics, how they were defined for the respondents and what scale was used to grade them. Some of the metrics have been used (sometimes under different names, discussed in Subsections 2.3.1 - 2.3.7) in previous research, but the subset used for this study is unique.

In addition to the 7 metrics, the respondents were asked to grade overall how satisfied they would have been when receiving such a counterfactual explanation. Satisfaction does not indicate a distinct value in the explanation but serves more as a benchmark for the other metrics and the explanation on the whole.

There were some noteworthy metrics omitted from the questionnaire. Recency is a value suggested to have importance by Byrne [3]. It was omitted due to its difficult implementation into counterfactual scenarios. For humans to evaluate a CE's recency, it either needs to be timestamped or all of the suggestions made need to be presented in a way where the passage of time can be deduced. Furthermore, Recency is not relevant in many counterfactual scenarios, such as recommending the user to get a raise. As the respondents need to grade every metric for every counterfactual, Recency would only be relevant where it was purposefully implemented and its interplay with the other metrics would be difficult to research. Other relevant omittals of Plausibility and External Coherence are discussed in Subsection 2.3.2 and the omittal of Trust in system in Subsection 2.3.4.

3.2 Creation of Questions for Questionnaire

After settling on the collection of metrics, 30 counterfactual scenarios that would sufficiently cover all dimensions of our chosen metrics were created. The data for the counterfactuals was taken from the Adult dataset [31] and the Pima Indians Diabetes dataset [32]. These datasets were chosen because of the variety of domains and variation of categorical and continuous data.

To generate this set of counterfactuals scenarios, Random Forest classifier models were trained on the Adult and Diabetes datasets with accuracies 85% and 92%, respectively. Next, a CSA DiCE [1] was used on dozens of factuals to generate hundreds of counterfactuals in order to choose situations exhibiting special cases of the mentioned metrics in Table 2. DiCE was chosen due to its in-built diversity in creating counterfactuals, offering a more varied set to hand-pick from.

The generated CEs were used for a basis for our counterfactual situations. To increase the variance of samples and to cover all dimensions of the metrics fully, some of the CEs in the questionnaire were modified with artificial information. For the same reason, the questionnaire also contains counterfactuals from fully artificial data.

To guide the generation of counterfactual examples, a set of dimensions that the metrics covered was created. Dimensions are unique ways a metric can manifest itself in a counterfactual situation. For example, in Feasibility's case, asking someone to earn 100 times more money is a different dimension of Feasibility than suggesting a change in their birthplace. Although both can be considered infeasible, one is suggesting a change in a continuous feature and the other in a categorical feature, which respondents could evaluate differently. Dimensions served as a guideline when creating counterfactual scenarios for the questionnaire. Covering all dimensions ensures that all facets of metrics are represented in the questionnaire, allowing for variability in the responses. The dimensions of Feasibility and Consistency are more nuanced, based on the separation of continuous and categorical data, while for the other metrics dimensions simply illustrate how well the explanations followed the decided definitions in Table 2. The dimensions are listed in Table 3. The questions referenced henceforth will be notated by their IDs (ID 5-34), referring to one of the questions found in Appendix I.

Table 3. Dimensions of metrics covered in questionnaire, and an example question from the questionnaire in Appendix I that covers that dimension

Metric	**Dimension**	**Question**
Feasibility	continuous value having an extreme out of distribution jump	ID 9
	continuous feature change from 0	ID 10
	continuous feature jump not from 0	ID 11
	continuous feature big jump within distribution	ID 12
	ordinal feature big change	ID 13
	ordinal feature change in wrong direction	ID 14
	nominal feature change	ID 15
	changing a non-actionable feature	ID 16
Consistency	two correlated continuous features changed in opposite directions	ID 17
	two correlated continuous features changed in mutual direction	ID 18
	two correlated categorical features changed in opposite directions	ID 19
	two correlated categorical features changed in mutual directions	ID 20
	two correlated features: continuous feature increased, categorical feature decreased	ID 21
	correlated continuous feature and categorical feature changed in mutual direction	ID 22
	two correlated features: continuous feature decreased, categorical feature increased	ID 23
Completeness	clearly incomplete explanation	ID 5
	relatively complete explanation	ID 24
	complete explanation	ID 30
Trust	scenario unlikely leads to the desired outcome	ID 5
	explanation doing some things towards the desired outcome	ID 24
	explanation likely leading to the desired outcome	ID 30
Understandability	explanation has good understandability	all*
Fairness	explanation changing a likely sensitive feature	ID 32
	explanation with no sensitive feature changes	ID 31
Complexity	explanation with one feature change	ID 5
	explanation with a couple feature changes	ID 6
	explanation with many feature changes	ID 34

All of the questions were made to have understandable text, as if the respondent does not understand the explanation, they cannot reliably grade the other metrics.

The dimensions listed in Table 3 focused more on bad examples, as humans inherently use negative information more [33]. Furthermore, all factuals used in the questionnaire had a negative outcome, for humans find it easier to change a situation from bad to good, as demonstrated by Byrne [3]. After formulating the questions, the questionnaire was made in the LimeSurvey environment [34]. In addition to grading counterfactuals, the questionnaire contained questions asking for simple demographic information:

- age;
- citizenship;
- highest completed level of education;
- level of English proficiency;
- whether they have experience with machine learning, counterfactuals, or medicine.

In order to carry out the questionnaire, permission from the University of Tartu Ethics committee was obtained. Before answering the questionnaire, the participants were made aware of how their data is processed, that they can opt out of answering the questionnaire at any time before the end of the survey and were asked to confirm, that they understand the terms.

3.3 Pilot Study

In order to test the questionnaire before gathering the bulk of responses, a pilot study with 29 questions was ran within the University of Tartu research groups. In total, 15 people filled out the survey. The main feedback and changes are gathered in Table 4.

Table 4. Feedback and changes made to the questionnaire during the pilot study.

Feedback	**Solution**
Metric "coherency" is vaguely defined.	Renamed coherency to Consistency and redefined Consistency to only account for consistency between parts of explanations.
The presented factuals and counterfactuals contain too much unnecessary and expository text. Example can be seen in Appendix I	Reduced redundancy and excessive wording in the instruction texts.

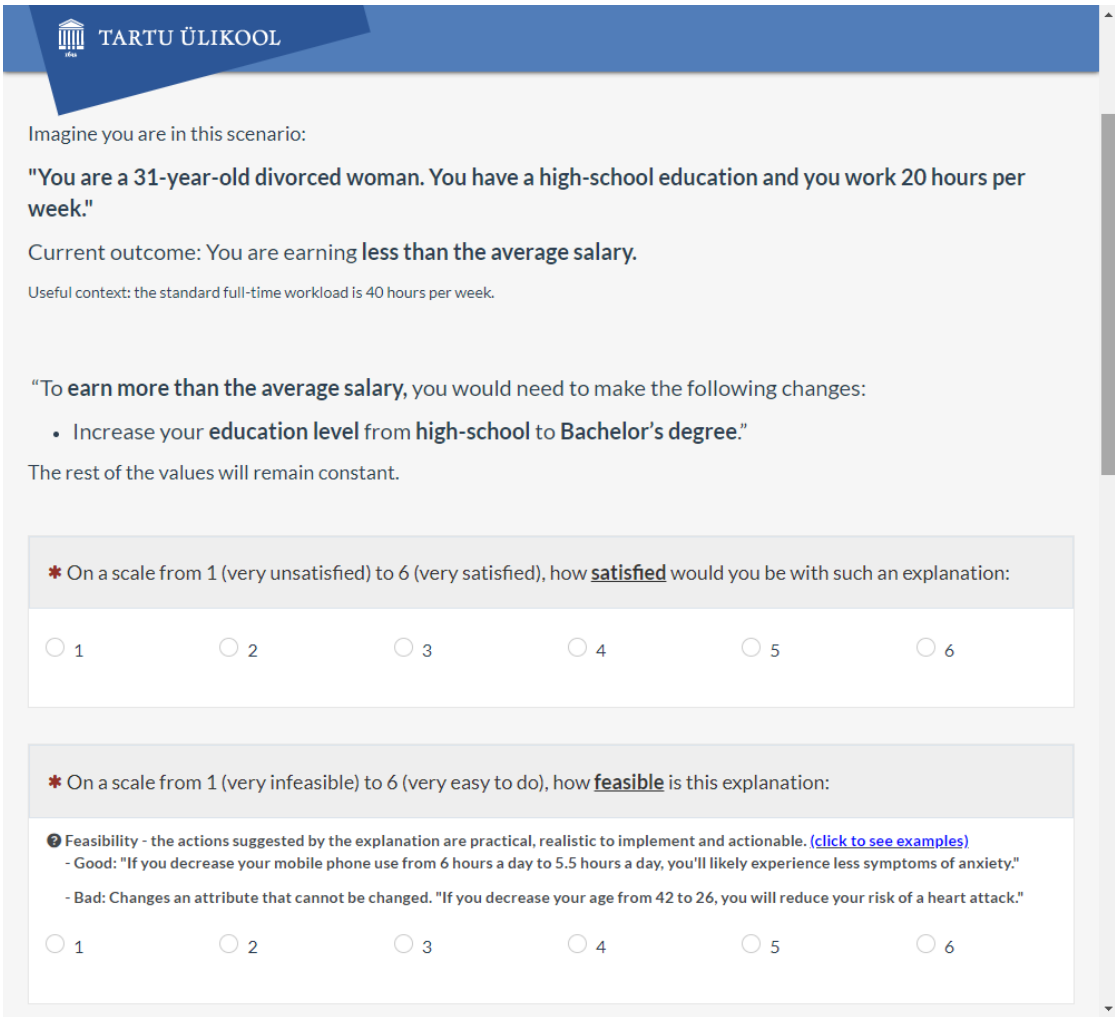
Unlike all other metrics, "bias" is a negative word and creates confusion, which direction is good in the Likert scale.	Changed bias name to Fairness.
Complexities scale is hard to understand in contrast to the others.	Changed Complexity's scale from 1...5 to -2...2 in order to make the difference between low and high Complexity ratings more obvious.

Additionally, counterfactuals were rarely rated to have high Complexity. To expand on the high-Complexity dimension a question with 6 feature changes (ID34) was added.

Generally the pilot study confirmed that the questionnaire covers both low and high extremes for all of the metrics, so no additional modifications were deemed necessary. Moreover, the preliminary correlations between Satisfaction and other metrics were strong, so the survey was moved to the second stage.

3.4 Gathering Answers from Questionnaire

Collecting responses for the survey was done through the online platform Prolific [35]. Prolific is a website, where people can post surveys and the respondents receive monetary compensation for answering them. Prolific was chosen due to its good reviews, rapid answer turnout and the possibility to restrict who can answer the questionnaire. Importantly, only people who are fluent in English gained access to survey it this study, as the language used in the questionnaire was complex. In total, 100 people were surveyed.


TARTU ÜLIKOOL

Imagine you are in this scenario:

"You are a 31-year-old divorced woman. You have a high-school education and you work 20 hours per week."

Current outcome: You are earning **less than the average salary.**

Useful context: the standard full-time workload is 40 hours per week.

"To **earn more than the average salary,** you would need to make the following changes:

- Increase your **education level** from **high-school** to **Bachelor's degree**."

The rest of the values will remain constant.

* On a scale from 1 (very unsatisfied) to 6 (very satisfied), how **satisfied** would you be with such an explanation:

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6

* On a scale from 1 (very infeasible) to 6 (very easy to do), how **feasible** is this explanation:

Feasibility - the actions suggested by the explanation are practical, realistic to implement and actionable. (click to see examples)
- Good: "If you decrease your mobile phone use from 6 hours a day to 5.5 hours a day, you'll likely experience less symptoms of anxiety."
- Bad: Changes an attribute that cannot be changed. "If you decrease your age from 42 to 26, you will reduce your risk of a heart attack."

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6

Figure 5. Example of question used in the LimeSurvey environment. The figure displays how the counterfactual situation was displayed to the respondents and how Satisfaction and Feasibility were asked to be graded, with definition and examples displayed for Feasibility. The 6 remaining metrics graded could be seen when scrolled down.

Figure 5 depicts the page where the respondents graded the counterfactual explanations. In the top of the page they could read the original factual situation and the corresponding counterfactual explanation. They then had to grade the CE in 8 categories. Before answering the questionnaire, the respondents were presented with the definitions of the metrics as listed in Table 2. In addition to the definitions, the respondents were provided with two basic counterfactual situations as examples for each metric (listed in Appendix I). One, where the expected score of the metric was good and one bad. The counterfactuals were presented to the respondents in a randomised order. For every metric there was an option to view how it was defined and one good and one bad example of that metric. To detect fraudulent respondents, an attention check was added into one of the counterfactual scenarios, which prompted the reader to write "here" into a text box in the survey, if they found that text.

The respondents were from a numerous countries and from across a varied educational background, documented in Table 5 and Table 6, respectively. 54 people answered that they have some experience in the field of machine learning. A surprisingly large amount of people had had contact with counterfactual explanations and/or causality frameworks with 27 respondents answering, that they are at least familiar with one of the terms.

Table 5. Citizenships of respondents

Country	**number of respondents**
Mexican	20
Portuguese	15
South African	12
Chilean	11
Polish	10
Greek	5
Other	27

Table 6. Education level of respondents

Degree	**number of respondents**
Bachelor’s degree or equivalent	48
High school	34
Master’s degree or equivalent	17
Primary school or lower	1

Generally, the respondents seemed to understand what they were evaluating, as an extra question asking how well they understood the definitions provided at the beginning of the questionnaire (Table 2) received an average score of 5.2/6.

4 Methods for Data Analysis

In order to gain insights about the meaningfulness and implications of the questionnaire, multiple data analysis methods were applied. Firstly, the data spread itself was explored to uncover significant patterns in the data. This was done using correlations, dimensionality reduction and clustering. Secondly, the metrics' relevance towards overall Satisfaction were studied using interpretable models to predict Satisfaction from the metrics.

4.1 Correlation and Statistical Significance

A very basic connection between metrics can be calculated through correlation. As the data gathered in the questionnaire is ordinal, Spearman is the appropriate correlation to use. Unlike Pearson correlation, Spearman first maps values into a ranking-based system and then calculates the correlation.

To assess whether the correlations calculated are statistically significant, a p-value test is run. In the case of correlation, p-value indicates the probability of getting a correlation as extreme or more with the assumption, that the features are uncorrelated. A correlation is statistically significant if it is smaller than a chosen significance level (α). The p-value test can also be run to test how significant the difference between two mean values are.

The widely used significance level is 0.05, but this thesis measures pairwise correlations between 8 different metrics (in total 28 correlations). When making so many calculations, the probability of randomly finding a statistically relevant correlation rises. To account for this, the Bonferroni Correction was applied on the significance level, such that $\alpha_{adjusted} = \alpha/n$, where n is the number of correlations calculated. Bonferroni Correction assures that the significant correlations discovered are not by random chance. In the case of this thesis $\alpha_{adjusted} = 0.05/28 \approx 0.00179$.

4.2 Dimensionality Reduction Methods

Principal component analysis (PCA) is a linear dimensionality reduction method originally developed by Karl Pearson [36] and formalised by Harold Hotelling [37]. PCA is used to form principal components out of multiple features, essentially compressing multiple values into fewer values. Principal components are constructed by finding a vector through the data, that maximises the variance of the projections on the vector cast by the distance from the data points. When generating multiple principal components, the new vectors need to be perpendicular to already found vectors. The data can then be cast along the found principal components, resulting in data with the same number of dimensions as the number of principal components. The first principal components are always more representative of the data, as the variance is always maximised and the next principal component cannot be correlated with any previous ones. PCA is especially useful when visualising high-dimensional data, as it can be compressed into 2 dimensions, which have the most representative values.

Factor analysis is another tool for dimensionality reduction. While PCA tries to explain the most variance within the data, factor analysis searches for underlying variables (called factors) able to cover the correlations between the true variables. The idea is that there may be fewer (originally unseen) features able to explain the originally measured variables. The survey asked the respondents to grade 7 metrics, factor analysis may expose that these 7 metrics can be be sufficiently explained by fewer combined metrics. PCA can also be used essentially as a tool for factor analysis. Namely, the variance of data that the components capture can be measured. These measurements can then be compared to see how many factors are needed, to cover most of the variance of the data.

Another dimensionality reduction method is **t-Distributed Stochastic Neighbor Embedding** (t-SNE), proposed by Maaten and Hinton [38]. Instead of linear operations, t-SNE calculates lower dimensions through a search function. Firstly, t-SNE calculates pairwise correlation between data points. Then for each point, it calculates the probability to pick any other point as its neighbour. After this, t-SNE attempts to find a lower-dimension space with the same amount of data points, where the probability distribution is similar to the one in the high-dimension data. As t-SNE is using gradient search to find the optimal distribution, it is not linear like PCA. An important variable for t-SNE is perplexity, which dictates how many neighbours a data point considers when formulating its probability distribution. Varying perplexity drastically modifies the structure of the output from large clusters to small.

In rough terms, PCA focuses on retaining the variability of data and t-SNE on retaining relationships between data points. For example, PCA can be used to visualise the variety of questions and t-SNE to find similarly answered questions in the questionnaire.

4.3 Clustering Questionnaire Respondents and Questions

Clustering can be used to visualise the correlations and connectedness of data. Density-based spatial clustering of applications with noise (DBSCAN) is an algorithm authored by Ester et al. [39] in 1996. As the name states, it clusters data on the basis of density. DBSCAN works with only two necessary parameters: eps (how close points should be to be clustered together) and minPoints (minimum number of points to form a dense region). The main advantage of DBSCAN is its robustness to noise, as points outside of dense areas do not get clustered. This robustness is especially useful when searching for outlier questions and/or respondents. Another strength is its versatility in discovering clusters, as it can deduce the number and shape of the clusters itself.

A multi-level clustering method is biclustering, first discussed by John Anthony Hartigan [40]. Biclustering does essentially what any other clustering method does, except along both axes' of a matrix or dataset. In terms of this thesis, this method allows the discovery of how specific respondents clustered for specific questions. A specialised biclustering method is spectral biclustering, which assumes that the used data has a checkerboard-like structure. This is especially useful when looking for rectangle-shaped patterns in data, which in this thesis's case represent a subgroup of respondents answering similarly for a subgroup of questions.

4.4 Machine Learning

To research how well the chosen metrics represent overall Satisfaction and if they can be used to predict it, machine learning models were used. It is a little counter-intuitive to use models in order to explain the importance of metrics designed to make explanations about the models themselves more understandable. However, some models are inherently built to be understandable and offer a good overview of the importance of metrics when predicting the overall quality of the explanation.

Machine learning models broadly split into two: classification and regression models. Classification models are designed to assign inputs into specific categories based on the patterns learned from the training data. These models are used for tasks where the outputs are categorical, such as determining whether a loan was approved or not, or identifying an email as spam. On the other hand, regression models predict a continuous value based on the input variables. These are suitable for tasks where the prediction involves quantifying something, such as estimating the amount of a loan given or predicting house prices.

An example of a model with clear, displayable argumentation is the **Decision Tree**. A Decision Tree is a model that uses a tree-like structure of decisions to process input data. The concept of using a branching tree to make decisions has been around longer than machine learning and has use even outside of computer science, such as medicine where doctors can use decision trees to diagnose diseases from symptoms. In a decision tree, each node examines a feature of the input and directs the data to the next node based on its value. This continues until the input reaches a leaf node, which provides the classification. The sequential nature of decision trees, examining one node at a time, makes it easy to trace the path from input to output. Consequently, when using an accurate decision tree with relevant features, the basis of decisions can be clearly understood.

A further development to the Decision Tree is the **Random Forest** (RF) model.Essentially, a random forest consists of multiple decision trees that process the input independently and then aggregate their outputs. In classification tasks, this aggregation involves counting votes from each tree to determine the final class, whereas in regression tasks, the model calculates the mean of all predictions from the trees. Although random forests generally perform better on larger datasets compared to single decision trees, their complexity can make them less interpretable [41], [42].

To test the hypothesis that predicting Satisfaction can be sufficiently achieved using only linear relationships between metrics, the **Linear Regression** model was employed. Linear Regression makes its predictions based on Equation 2. By adjusting the coefficient vector, Linear Regression aims to find a "line" through the data that minimizes the sum of all squared distances between this line and the target values. This method essentially attempts to create the best linear fit to explain the observed relationships.

$$
\begin{aligned}
&y = X\beta + b \\
&\qquad y = \text{Target (Satisfaction)} \\
&\qquad X = \text{Feature matrix} \\
&\qquad \beta = \text{Coefficient vector} \\
&\qquad b = \text{Intercept}
\end{aligned}
\tag{2}
$$

When models are built on imbalanced data, they are themselves imbalanced and biased. There are multiple **resampling** methods to combat this. Random undersampling drops random data points with the superfluous class. Random oversampling generates copies of random data points with the less common class. SMOTE-TOMEK, a resampling method proposed by Gustavo et al. [43] is a combination of the oversampling method SMOTE developed by Chawla et al. [44] and the undersampling method TOMEK based on the research by Ivan Tomek [45]. SMOTE-TOMEK first generates new synthetic samples that are close to the minority class based on Euclidean distance. Then it eliminates data points from the majority class, for which the closest neighbour and vice-versa is a data point from another class. This method creates new data while distancing classes from each other. As the questionnaire has more bad CEs, applying this approach on the response data would likely make the decision space for well-rated instances denser, while also eliminating outliers among poorly rated counterfactuals that are situated close to highly rated ones.

To ensure that the model is trained on and validated on the highest number of data points, **cross-validation** can be used. Cross-validation splits the training data into multiple sets, giving the possibility to variate the validation set used during training. This ensures that the trained model has been tested with as much data as possible and that it does not overfit for only one subset.

5 Results

After gathering the results to the survey, data analysis methods described in Section 4 were applied on the data. Firstly, data cleaning was applied to get rid of suspicious responses. Secondly, the underlying data distribution and connections were explored. Finally, machine learning models to predict Satisfaction from the metrics were built with the aim of finding out, what metrics are most useful.

5.1 Data Cleaning

In order to verify the data received from Prolific, multiple filtering methods were applied to detect suspicious respondents. Potentially fraudulent respondents were detected through plotting respondents with reduced dimensions, utilising both t-SNE and PCA. Respondents were visualised both on the level of the whole questionnaire and the level of answers to singular questions. When the same respondent id-s appeared as outliers multiple times, they were marked down in a table gathering statistics potentially indicating fraud (Table 7). An example of how DBSCAN found outliers in data reduced to 2 dimensions with PCA can be seen in Figure 6.

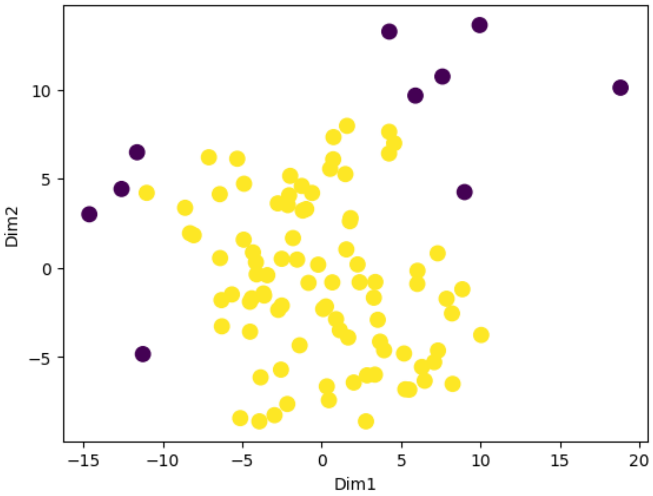


Figure 6. Detecting outliers by using DBSCAN on respondents' answers.

Table 7. Data cleaning summary table, listing suspicious respondents and how suspiciousness was determined. Respondents who were suspicious in 3 or more categories were dropped from the study. Legend: x - suspicious, ? - mildly suspicious, R - remove

Respondent id	*Survey filling time*	*Failed attention check*	*PCA suspicion*	*t-SNE suspicion*	*Similar answer pattern*	*Suspicion question 16*	*Suspicion question 21*	*Suspicion question 32*	*Avg understandability < 3*	*Conclusion*
64			x			?				
66	?	?								
72		x								
78		x	x		x	x	x	x		R
91						?		?		
92										
95	?	x	x							
96	x	x	x		x		x	?		R
98					x	x	x			R
114		x			x	x	x			R
146		x			x		?			
147					x			x		
148						?			x	R
159		x	?		x	x	x	?	x	R
163			x	x	x		?			R
170			?		x	x	x	x	x	R
172					x	x				
182			?		x	x	x	x	x	R

Secondly, all the respondents answer vectors were clustered into 30 clusters (chosen based on the total amount of questions). If a single respondents' answer vectors were clustered into the same cluster at least 10 times, or into 2 clusters a total of 15 times, the respondents' id was noted in Table 7. This check detects respondents who consistently rated the questions similarly, possibly indicating mindless grading.

Thirdly, the respondents who failed the attention check and had suspiciously low survey completion times were marked. Those whose average Understandability score was below 3 throughout the survey were also marked, as this either indicates a lack of understanding of the metric or the text being evaluated. Either case is unacceptable and does not provide worthwhile data.

Next, the identified suspicious respondents were individually analyzed in the context of specific indicator questions. For example, question 16 (ID16) recommends that the reader change their country of birth. Respondents who rated this counterfactual with high Feasibility were flagged, because changing one's place of birth is not feasible under any circumstances.

Finally, based on the information gathered in Table 7, 9 out of 100 total participants were removed from the study for failing three or more checks. Additionally, it was verified that the Prolific respondents did not answer completely differently from the pilot study respondents through clustering them together, detailed in Appendix III

5.2 Patterns in Response Data

In order to verify that the questionnaire sufficiently covers what humans look for in counterfactuals and that the biases were correctly selected, the connections between responses were studied. Firstly, whether the questions themselves are graded diversely was analyzed through clustering. Secondly, how the respondents graded the metrics in theorised points of importance (marker questions) was studied. Finally, the relevancy of each metric was tested through correlations and factor analysis.

5.2.1 Statistics and Variance of Questions

To assess how comprehensively the questions in the questionnaire covered all metrics and to evaluate the overall diversity of the questions, statistics were calculated for the mean metric values of each question and amassed in Table 8.

Table 8. Summary table for metric statistics calculated by taking mean of each statistic calculated individually for each question. Coefficient of variance is N/A for Complexity, as its scale does not have a meaningful zero.

	mean	Stdv	Variance	Coef. Of var	Min	Max
Satisfaction	2.9	1.27	1.67	0.44	1.33	5.19
Feasibility	3.17	1.27	1.68	0.4	1.32	5.07
Consistency	3.63	1.41	2.03	0.39	1.7	5.49
Completeness	3.27	1.43	2.06	0.44	1.71	5.32
Trust	3.09	1.35	1.9	0.44	1.36	5.33
Understandability	4.85	1.33	1.91	0.27	4.03	5.53
Fairness	3.92	1.54	2.44	0.39	1.52	5.4
Complexity	-0.33	1.07	1.19	N/A	-1.09	0.9

From the min and max values in Table 8, we can say that the survey covered all extreme dimensions of the gradable metrics successfully. The only extremes not addressed in our questions were low Understandability and very high Complexity. Even in the case of a counterfactual with 6 feature changes (ID34), respondents rated it on average with a Complexity of 0.9. The omission of low Understandability was intentional, as respondents who do not understand the text cannot reliably grade other metrics.

To visualise the questions' relations to each other and to find potentially interesting outlier questions, the questions' 7 average metric values were reduced to 2 dimensions using t-SNE and visualised in Figure 7. t-SNE was used in this case, as it retains the proximity of similar questions.

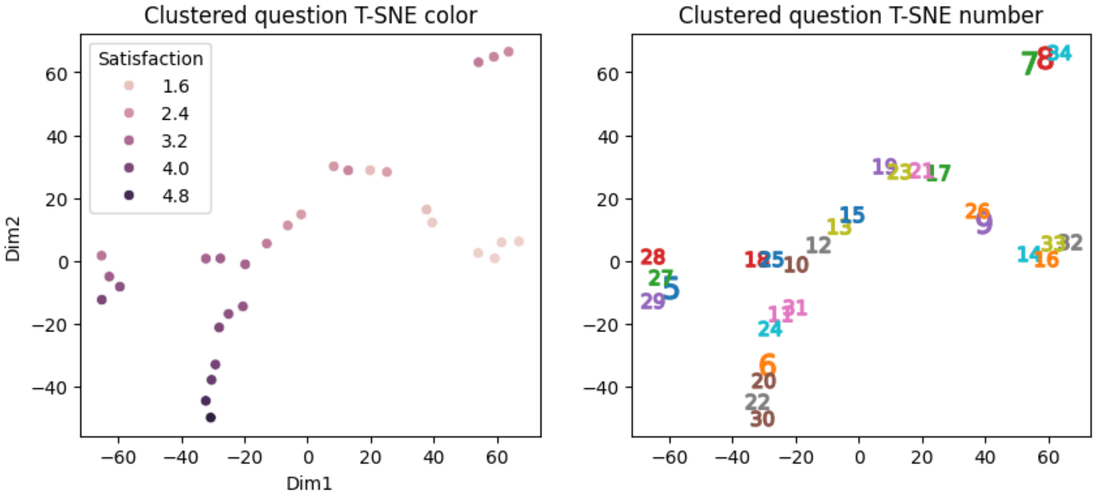


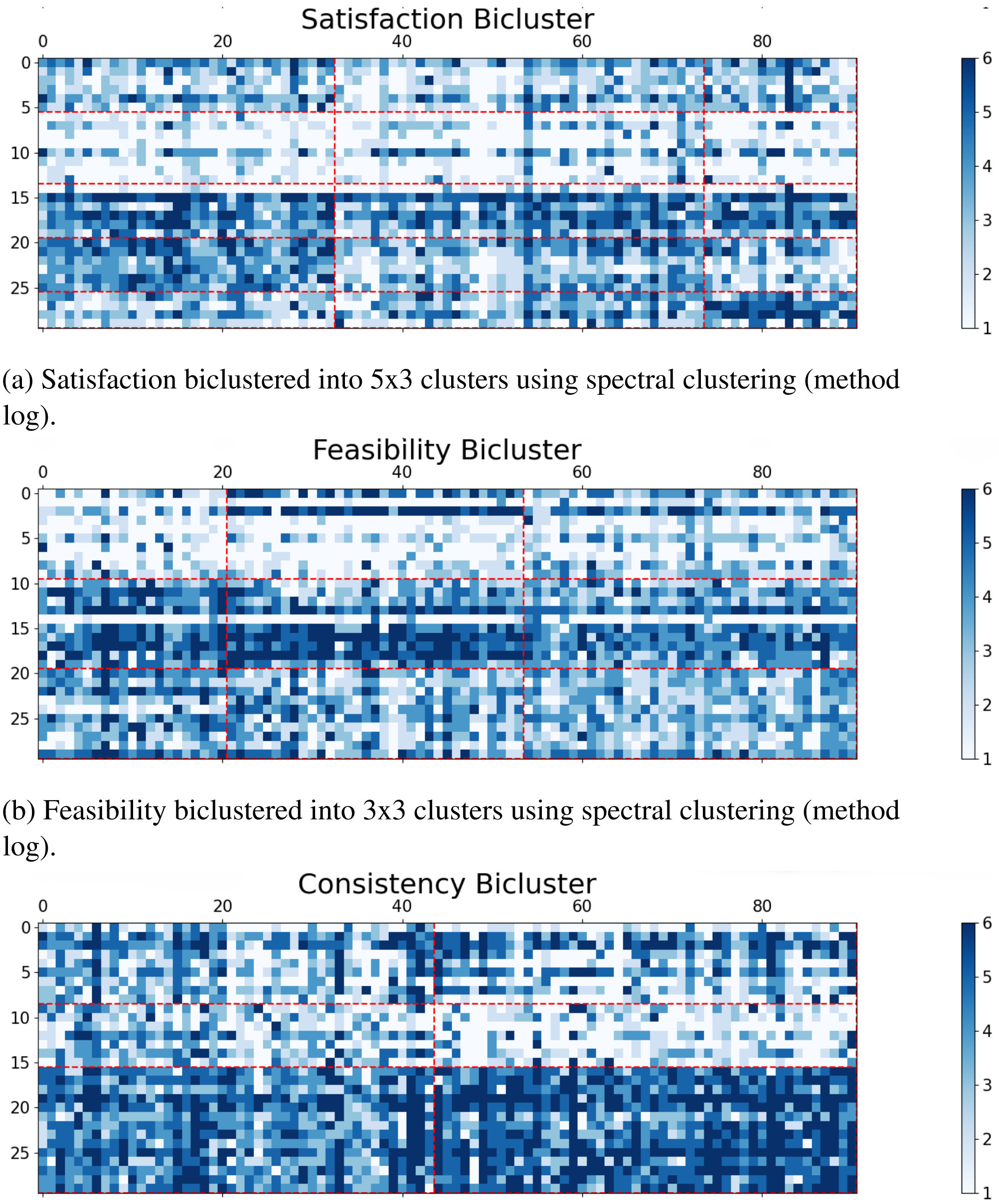
Figure 7. Questionnaire questions' 7 average metric values reduced to 2 dimensions (t-SNE perplexity 3).In the left image the questions are colored by average Satisfaction, in the right the corresponding question id-s

The left image of Figure 7 is colored by the average Satisfaction rating of the question ranging from 1-6 and the right image is numbered by the corresponding question ID. Due to the questions in Figure 7 only accounting 3 closest neighbours (perplexity 3) through the t-SNE process, a continuous chain forms, going through all levels of Satisfaction. This illustrates the questionnaire's coverage of the whole counterfactual quality spectrum. When clustering more generally with all questions accounting for 25 closest neighbours, a more general distinction forms, as seen in Appendix III. Notably, on Figure 7 a subset of questions 7, 8, 34 and 5, 27, 28, 29 are separate from the general chain. These questions stand out due to their high Complexity scores and low Complexity scores, respectively.

5.2.2 Metric Representation in Questionnaire

In order to assess the coverage of metrics in the questionnaire and identify any notable trends or outliers in perceptions, biclustering was employed. This technique is particularly effective for identifying subsets of respondents who answered a specific subset of questions similarly, potentially revealing groups with distinct biases. The choice of the number of horizontal and vertical clusters, as well as the normalization method, was determined through visual inspection, to identify which combinations best represented the data. Additionally, the questions that targeted the planned dimensions listed in Table 3 were analysed to evaluate how closely the respondents' ratings matched expected outcomes. For easier comparison, the questions were ordered by average metric rating and average standard deviation for each metric separately, as shown in Appendix III

(Table III.1 and Table III.2, respectively). Detailed analysis of each metric is described in the following section.



(a) Satisfaction biclustered into 5x3 clusters using spectral clustering (method log).

(b) Feasibility biclustered into 3x3 clusters using spectral clustering (method log).

(c) Consistency biclustered into 3x2 clusters using spectral clustering (method log).

Figure 8. Biclustering applied on the answers of the questionnaire. Rows represent (30) questions and columns (91) respondents. Each cell value reflects how one person rated one question in one metric. Red dotted lines separate the edges of the clusters

Satisfaction did not cluster well, and no truly distinct clusters formed, as seen in Figure 8a This outcome is understandable since the questions were originally designed to elicit a variety of rating.

The bicluster (Figure 8b) depicting **Feasibility** clearly splits into three levels: low Feasibility, where respondents likely considered the changes impossible, medium Feasibility, which shows considerable variation from person to person, and high Feasibility. The questions from row 20 to 30 in the Figure 8b are especially important, as there appears to be no consensus among the respondents' clusters. While most people agree on whether an explanation is either very unfeasible or very feasible, preferences become less clear in the middle range.

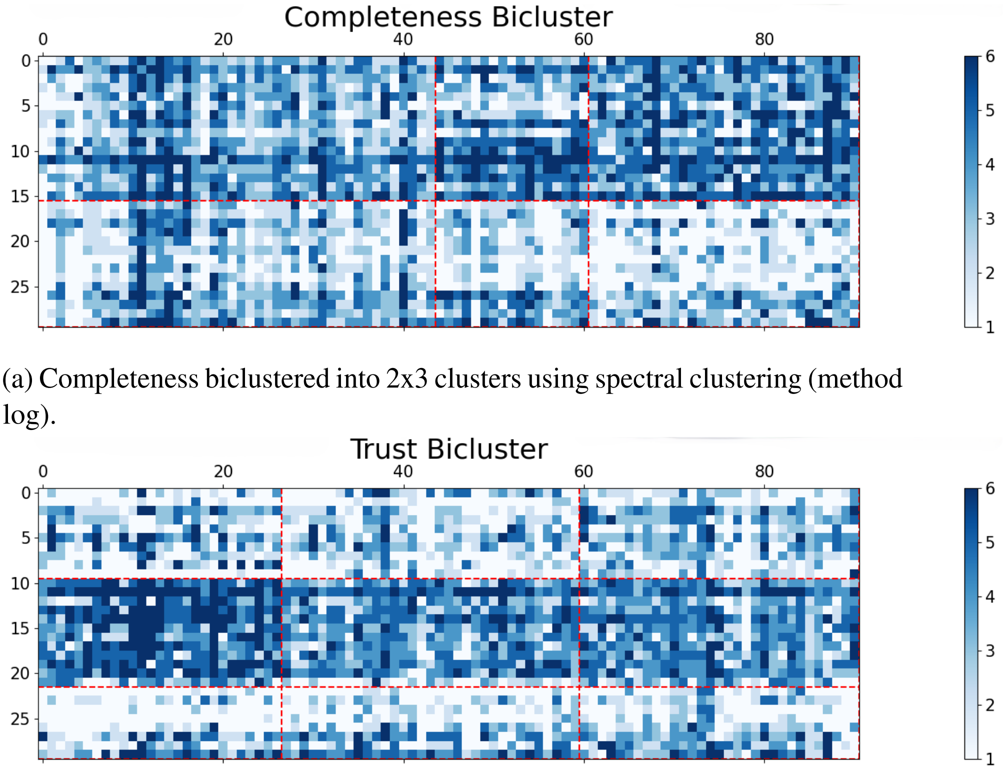
The counterfactuals with the lowest rated Feasibility ratings involved changes to likely non-actionable features, such as gender and age. In contrast, counterfactuals designed to examine reactions to changes in continuous features did not perform as badly, though they still fell below the average. The exception being the largest recommended jump from 20 to 95 work hours per week (ID9), which was the second worst-rated on Feasibility.

Interestingly, changes in some features were considered more infeasible than others. For instance, the lowest Feasibility rating (ID32) was for changing gender from male to female. At the same time, changing one's country of birth (ID16) was only the sixth worst-rated counterfactual, with a change in work position from junior researcher to lawyer (ID26) considered as more infeasible. Moreover, a change from barista to hairdresser (ID15) was rated noticeably higher in Feasibility than both previous examples. This indicates that among nominal features, some changes are detected to be more feasible than others.

The respondents detected an average-to-low difference between changing a feature the same amount from 0 or from an already established number. For instance, recommendations to increase passive income from 0 to 1500$ per month and from 1500$ to 3000$ per month (ID10, ID11) were rated on average to have a feasibility of 3.09 and 3.42, respectively.

Regarding **Consistency**, there is a stark difference between two groups of respondents. On Figure 8c, the group from column index 44 onward have a more black and white view, while the rest have a more balanced view towards both high-Consistency and low-Consistency counterfactual explanations.

For Consistency, the worst rated counterfactuals were those originally targeted, as listed in Table 3. Interestingly, although many of these examples were structured similarly, their Consistency values varied. This variation stemmed from differences in the data and the types of data involved. For example, in counterfactuals where the attributes modified were hours studied per week and average grade in order to get into a university (ID21, ID23), the case where hours studied was decreased while the grade was increased received a significantly better rating for Consistency (p-value = 0.0024) compared to the case with the opposite changes. Although these scenarios are essentially identical in terms of what they test for Consistency, some respondents may have let their sense of the overall quality of the explanation to influence their judgement of Consistency. Notably, increasing the grade was deemed more crucial towards getting into university (ID23 Satisfaction 2.65 vs ID21 Satisfaction 1.69). How a person perceives the value of features may leak into their rating of Consistency.



(a) Completeness biclustered into 2x3 clusters using spectral clustering (method log).

(b) Trust biclustered into 3x3 clusters using spectral clustering (method bistochastic).

Figure 9. Biclustering applied on the answers of the questionnaire. Rows represent (30) questions and columns (91) respondents. Each cell value reflects how one person rated one question in one metric. Red dotted lines separate the edges of the clusters

As seen in Figure 9a, **Completeness** largely clustered into two categories: an explanation was either not complete at all, or it was somewhat complete. Explanations rated with very high Completeness were rare. The highest Completeness reached (ID30) was 5.32, and from there was an immediate drop-off to 4.87 (ID22).

The counterfactuals rated the worst were not the ones deliberately placed in the questionnaire, but were characterized by the same attributes. Namely, explanations containing only one feature change and not directly connected to the outcome. The lowest Completeness score was attributed to counterfactuals that had other bad traits about them as well, such as poor Feasibility or Trust.

Interestingly, when one-feature-change counterfactuals directed a change in a feasible way, they were not rated as badly. For example, when recommending ways to earn more than the average salary, the recommendation to earn a Bachelor's degree (ID5) was rated to have a significantly higher Completeness (p-value < 0.0001) than recommending to increase work hours from 20 to 95. Respondents either felt that earning a higher degree is a more effective way towards earning money than working more hours, or let the overall quality of the explanation affect their rating.

In Figure 9b **Trust** clusters into two main groups: good and bad. This may be due to the seemingly binary essence of Trust: a person either believes that following the explanation would fix the problem or it would not.

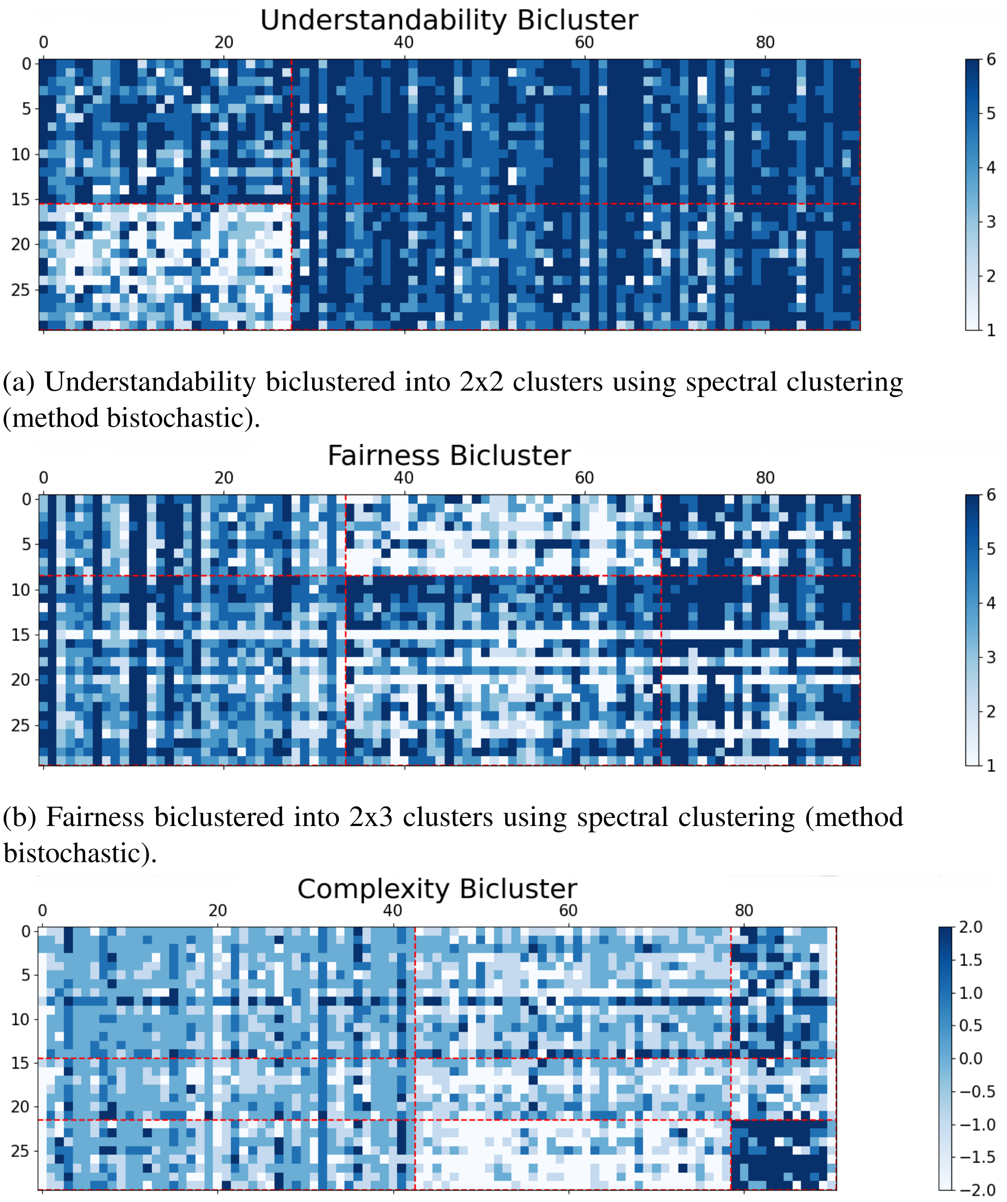
Trust, like Consistency and Completeness before it, seems to have a relation with Feasibility and the overall quality of an explanation. In order to earn more than the average salary, a counterfactual instructing to get a Bachelor's degree had a Trust rating of 4.0 (ID5). At the same time, a counterfactual seeking to achieve the same thing recommending five (positive or not related) feature changes with earning a Bachelor's degree among those changes (ID8) received a Trust rating of 2.74. This is a significantly lower (p-value < 0.0001) rating than the first, while doing more towards reaching the goal. This shows a clear bias towards every other metric seeming bad when Feasibility is bad or many of the respondents not understanding the definition of Trust.

Respondents evaluating **Understandability** were visually split into two groups in Table 10a: those who understood everything and those who did not understand a certain subset of questions. Interestingly, those who did not understand that subset also reported lower understanding of other questions.

The questions that formed the low-Understandability quadrant typically involved a feature change where the path to achieving the change was unclear. For example, the explanation with the lowest Understandability recommends changing education level from PhD to Master's (ID14). Respondents might have falsely interpreted Understandability to mean understanding how to fulfil the instructions given, potentially mixing it with Trust. In the case of reducing education level, they might have been confused about whether the explanation was suggesting them to expose themselves of plagiarism to lower their degree.

Out of all the biclusters, **Fairness** one has the most vertical structure, as seen in Figure 10b. This may reflect the individual differences in perceptions of what is right. One of the respondents even marked everything to be maximally fair and while unconventional, this opinion cannot be disregarded.

The lowest Fairness was received by the explanations that changed gender (ID32), country of birth (ID16), or relationship status (ID33). In total, only 8 counterfactuals were rated below the scale's midpoint of 3.5 for Fairness.



(a) Understandability biclustered into 2x2 clusters using spectral clustering (method bistochastic).

(b) Fairness biclustered into 2x3 clusters using spectral clustering (method bistochastic).

(c) Complexity biclustered into 3x3 clusters using spectral clustering (method log).

Figure 10. Biclustering applied on the answers of the questionnaire. Rows represent (30) questions and columns (91) respondents. Each cell value reflects how one person rated one question in one metric. Red dotted lines separate the edges of the clusters

Notably, counterfactuals where the factual situation described a person in a more sensitive situation had the highest standard deviations in Fairness, with a single woman with child having the top 2 spots. This could indicate some people also considering the original situation to determine if a counterfactual is fair. If that is the case, integrating Fairness into a counterfactual algorithm might need to account for processing the factual as well.

As seen in Table 10c, **Complexity** displays a relatively uniform bicluster, with most values clustering near perfect Complexity. An abnormal cluster is located in the rows 22-30. Respondents had differing opinions of complexity regarding this subsection of questions. Namely, the counterfactuals in this group were rated to be to simple, too complex and perfectly complex. All counterfactuals in this cluster involve infeasible changes. Respondents in the high-Complexity cluster might have considered explanations that make difficult changes to be complex in terms of "difficulty to execute".

Overall, respondents did not find many counterfactuals to be too complex as only four were rated to have higher Complexity than the ideal. Each of these counterfactuals had a Sparsity (number of feature changes) of at least 3. As expected, Sparsity was positively correlated with Complexity ($\rho = 0.31$, p-value < 0.00001). Sparsity was also positively correlated with Satisfaction ($\rho = 0.16$, p-value < 0.00001). However, this correlation is small and does not definitively support the view that shorter explanations are preferred, as suggested by previous studies [25], [26] nor does it confirm findings to the contrary [21].

5.2.3 Metric Correlations and Factor Analysis

To explore how interconnected the metrics were with each other and with Satisfaction, Spearman correlation was calculated between all individual scores and gathered in Figure 11. Additionally, the mean metric values of questions were plotted pairwise in Figure 12, providing a visual analysis of how the dimensions of different metrics interrelate, accompanied by metric histograms.

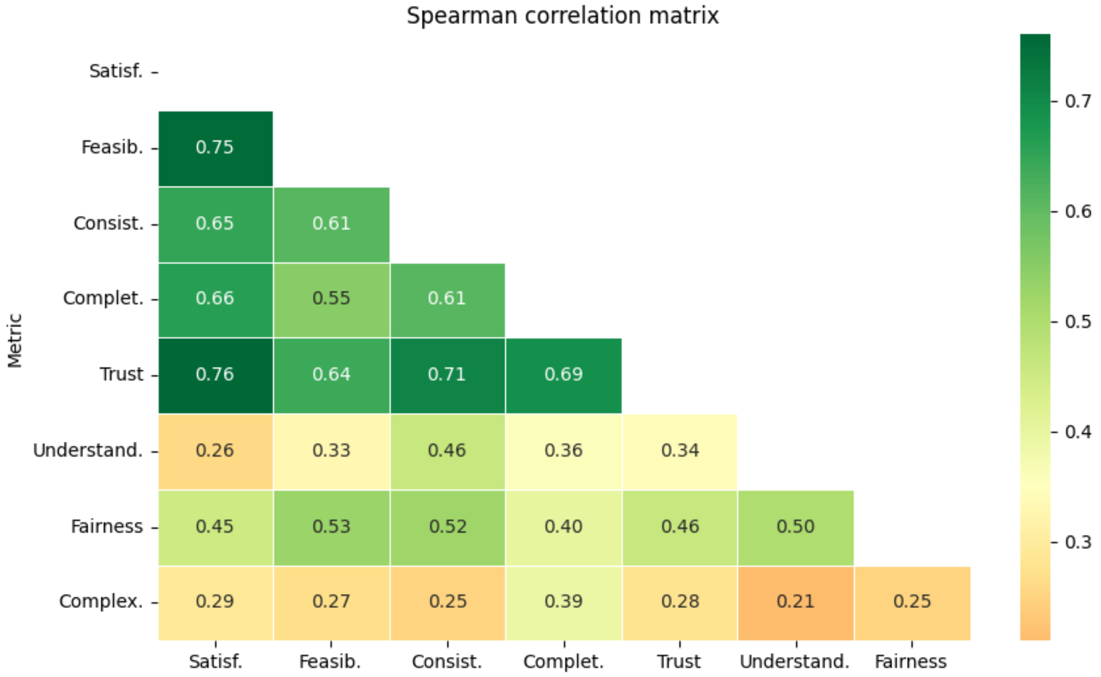


Figure 11. Spearman correlation table between metrics. For Complexity, the values were mapped to their absolute value and the order reversed, to get a scale from 0 to 2, where 2 is best and 0 bad Complexity in either direction. For all of the correlations, p-value < 0.00001, $\alpha = 0.00179$

It is evident from Figure 11, that most metrics are significantly correlated with Satisfaction and each other. Notably, Satisfaction, Feasibility, Consistency, Completeness and Trust are all heavily correlated with one another. While correlations among metrics and Satisfaction are beneficial, correlations among the metrics themselves could indicate underlying factors influencing the measurements.

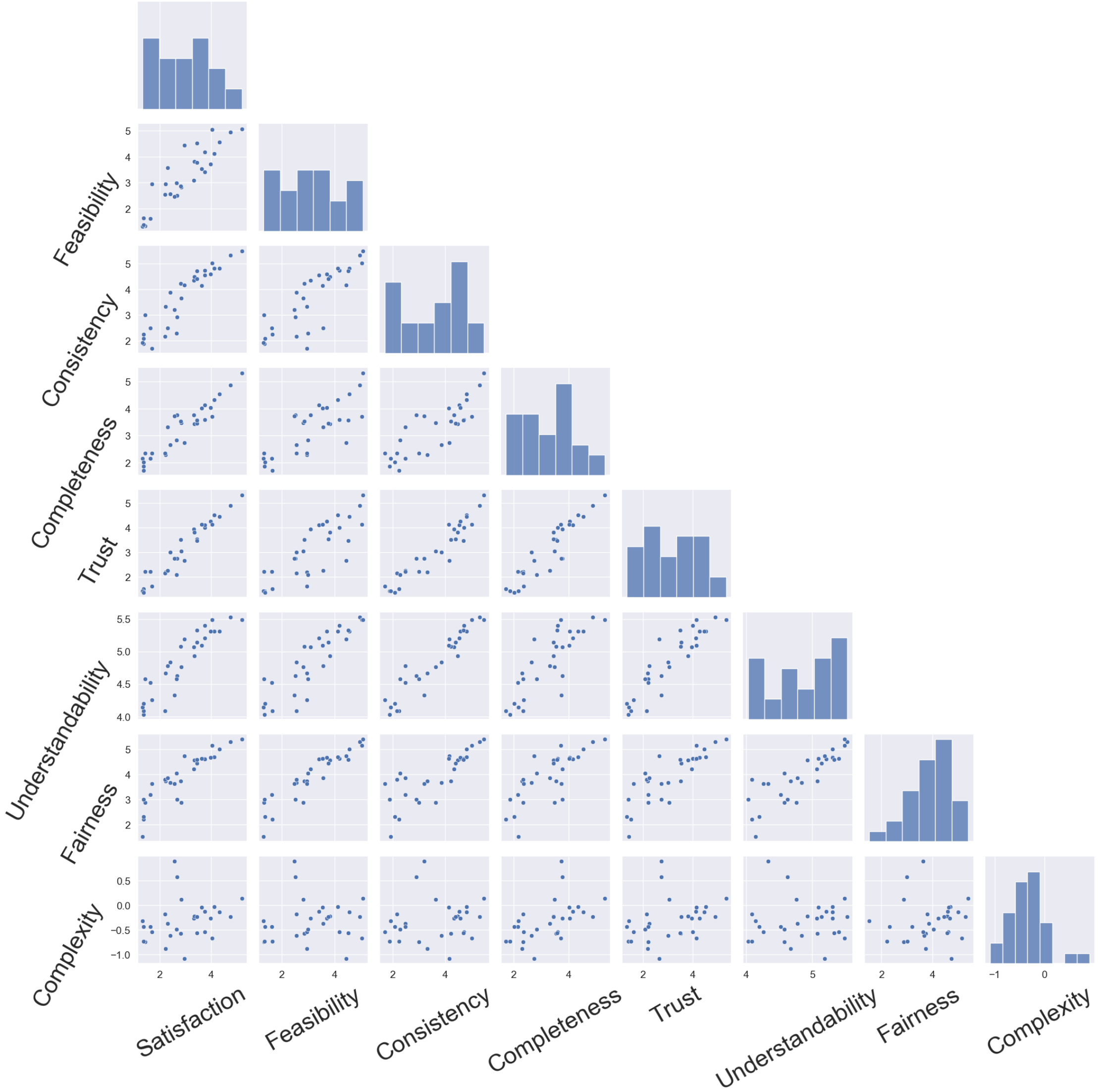


Figure 12. Pairwise scatterplots between all metrics. Each point on each subplot accounts for one question. The diagonal displays the histogram of each metrics through questions.

Figure 12 further illustrates how similarly the metrics were rated pairwise. Nearly all of the plots display a line of questions from badly rated to good. A good indicator that the questions cover the metric scales well is that the lines are mostly uniform. An outlier is Complexity, as it is measured on a different scale and mostly shows ratings clustering around perfect complexity with a few outliers.

These high correlations could point to a similar situation as when the Big Five personality traits were discovered. Originally, 37 personality factors were researched but the number was reduced to 5 through factor analysis [46]. To test this, factor analysis was run on 7 metrics (without Satisfaction).

As seen in Figure 13, the way a person rates the metrics can largely be explained by a single factor. Only two factors have an eigenvalue over 1 (meaning it covers more variance then a single variable). The elbow point occurs around the third factor.

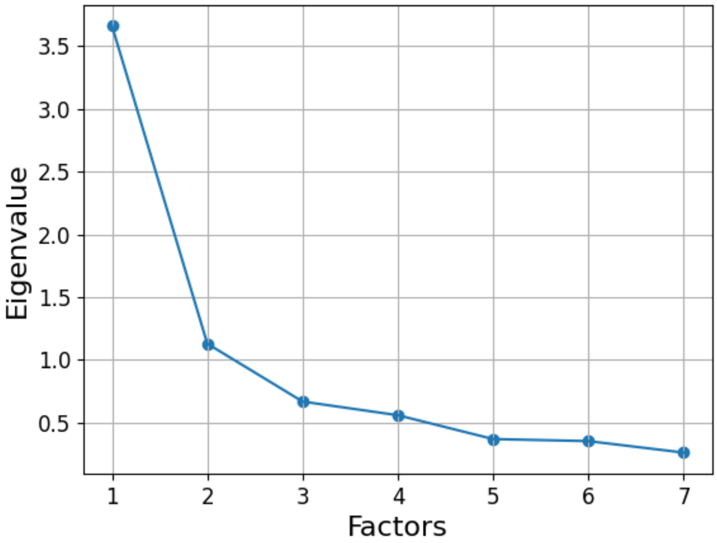


Figure 13. Scree plot depicting factor eigenvalues from factor analysis carried out on 7 metrics

Table 9. First 3 loadings of factor analysis used on 7 metrics

Factor	Feasibil.	Consist.	Complet.	Trust	Understand.	Fairness	Complex.
0	0.761	0.737	0.551	0.736	0.335	0.632	0.035
1	0.132	0.256	0.255	0.167	0.938	0.3	-0.057
2	0.153	0.215	0.623	0.403	-0.05	-0.045	0.38

Table 10. PCA components when reducing 7 metrics into 3 dimensions. Factor 0 is the most expressive and so forth.

Factor	Feasibil.	Consist.	Complet.	Trust	Understand.	Fairness	Complex.
0	-0.42	-0.47	-0.40	-0.46	-0.26	-0.40	-0.05
1	-0.01	-0.08	-0.41	-0.32	0.39	0.67	-0.35
2	-0.53	0.35	0.14	0.02	0.60	-0.38	-0.29

However, both Factor 0-s in Table 9 and Table 10 significantly influence five of the metrics. This suggests that the most expressive factor correlates highly with five metrics. To isolate this factor, the concepts of Feasibility, Consistency, Completeness, Trust, and Fairness would need to be unified under a single concept. In other words, an explanation fulfils this factor if it "makes sense". However, this definition does not contribute directly to the development of a human-preferred counterfactual algorithm. Another possibility is that the respondents' general opinion of the rated counterfactuals affected their views on individual metrics.

For a more visual overview of how metric ratings varied across the questionnaire, a line graph depicting the metric values through all of the questions was created (Figure 14).

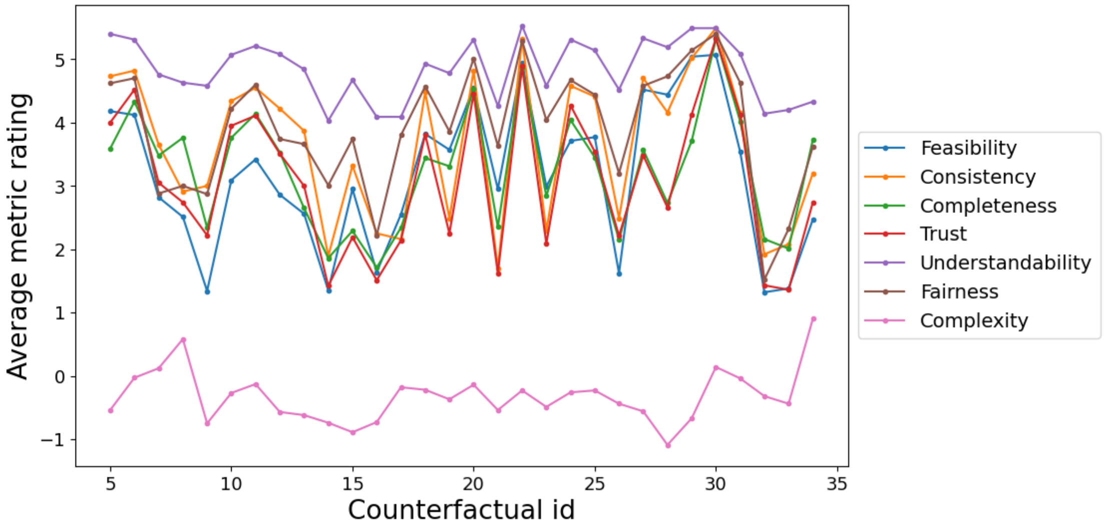


Figure 14. Line graph depicting the average scores of the 7 metrics through questions 5-34

From Figure 14, it is particularly evident that the mean metric predictions for individual counterfactuals not only trend similarly but also fall within a similar value range.

To assess person-by-person differences more granularly and determine which metric was most closely linked to Satisfaction for each individual, correlations between Satisfaction and other metrics were calculated separately for every person and for each metric in Table 11.

Table 11. Number of persons with the corresponding metric having the largest correlation with satisfaction

Metric	**Number of largest correlations**
Trust	38
Feasibility	27
Completeness	9
Consistency	8
Fairness	8
Understandability	1
Complexity	0

As shown in Table 11, different individuals identified different metrics as their primary predictor of Satisfaction, suggesting that personal preferences play a significant role in determining what aspects of explanations are most important to people. However, Trust and Feasibility emerged as the most important for the majority.

5.3 Predicting Satisfaction From the Measured Metrics

The high intercorrelations among metrics raise the question: what metric or combination of metrics is required to predict the overall Satisfaction with an explanation. To solve this problem, multiple machine learning models were trained to determine which metrics are most predictive of Satisfaction. The resulting dataset contained 2730 data points, covering 30 questions answered by 91 respondents. The data was split into three subsets: 1830 data points (61 respondents) for training, 450 (15 respondents) for validation and 450 (15 respondents) for testing. The data was split by respondents, as this guarantees an equal amount of representation for each counterfactual scenario in all sets. The split of respondents into subset was random.

5.3.1 Regression Model to Predict Satisfaction

Regression models offer a general indication of how well it is possible to estimate Satisfaction as a real number from the 7 metrics. Firstly, in order to determine if this problem can be solved with a linear solution, a linear regression model was used. The absolute error and root mean squared error (RMSE) of the model on the test set were 0.65 and 0.89, respectively. Considering that the standard deviation of Satisfaction in the questionnaire was 1.27, the model outperforms random guessing. The feature coefficients from the Linear Regression model are presented in Table 12. Feasibility and Trust were found to have the most significant linear influence on Satisfaction. The negative coefficient of Understandability might be due absence of badly worded explanations, which could misadjust the direction of influence.

Table 12. Feature coefficients in Linear Regression model.

Feasibility	**Consistency**	**Complet.**	**Trust**	**Understand.**	**Fairness**	**Complexity**
1.974	0.498	0.688	1.603	-0.434	0.079	0.295

Another model used for testing was Random Forest Regressor, as it is known to be a powerful model. Both fixed validation set training and cross-validation were employed, with a parameter sweep was run on both of the models, specified in Appendix IV. The best model achieved a mean absolute absolute error of 0.57 and RMSE of 0.8 on the test set, marking a slight improvement over the linear model.

5.3.2 Classification Model to Predict Satisfaction

To trace a clear path of "thinking" from input to Satisfaction grade, a Decision Tree model was trained. To reduce the complexity of a problem, classification was employed instead of regression. More importantly, the critical aspect of this model is not its accuracy, but its interpretability and the insights it provides into decision-making processes. The responses, based on a 6-point Likert scale, were categorised into three classes:

- 1,2 : low;
- 3,4 : medium (mid);
- 5,6 : high.

Given that the data from the questionnaire showed a heavy bias towards lower Satisfaction scores, multiple datasets were generated using undersampling, oversampling, and SMOTE-TOMEK to create the most representative model. Both a parameter sweep and cross-validation were conducted on each of these datasets, as detailed in Appendix IV.

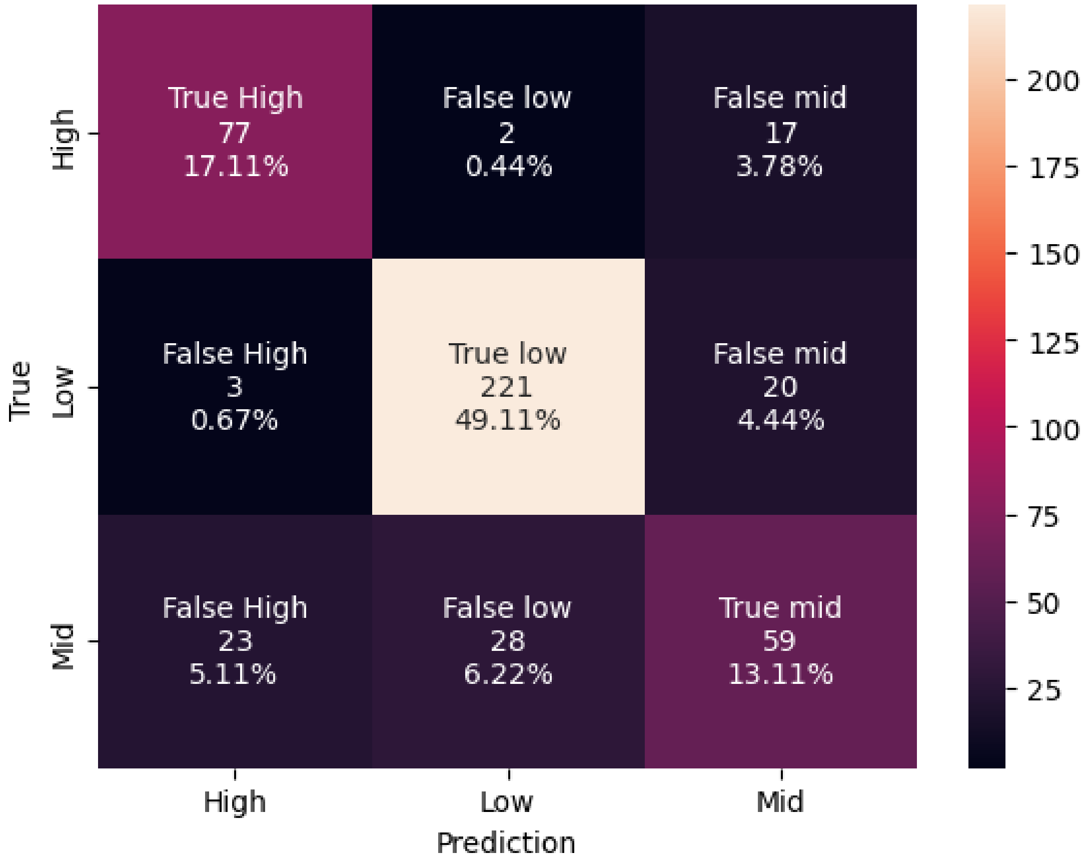


Figure 15. Confusion matrix for a Decision Tree classifier model trained with cross-validation on data resampled with SMOTE-TOMEK.

All of the models trained had an accuracy around 80%. Given that randomly picking between three classes in a balanced dataset would yield an accuracy of approximately 33%, the models demonstrated good accuracy. As shown in Figure 15, high and low classes were only misclassified into the opposite class 5 times in total, meaning the model had a good sense of quality-direction. The amount of times answers about specific counterfactuals were falsely predicted ranged from 0 to 6 (out of 15), and the amount of times answers given from specific individuals were misclassified ranged from 2 to 11 (out of 30), as seen in Appendix IV (Table IV.3 and Table IV.4, respectively). The high variance in misclassified data points between individuals may indicate either the models learned the preferences of certain individuals better or that some respondents frequently answered in uncommon ways.

The decision-tree models trained were largely dependent on only 2 metrics: Trust and Feasibility. To verify this, 25 models with at least accuracy 75% were trained on random subsets of the training set and the feature importances for prediction were extracted.

Table 13. Average feature importance of a decsisiontree classifier over 25 trained models.

Metric	Average feature importance
Feasibility	0.453
Trust	0.417
Completeness	0.072
Consistency	0.026
Complexity	0.013
Fairness	0.011
Understandability	0.008

As seen in Table 13, Feasibility and Trust are the dominant factors in the decision making process. An example of a model where Trust and Feasibility were immediately used to separate all low values can be seen in Figure 16. In that example, only Completeness was additionally used, determining whether an explanation achieves high or medium Satisfaction.

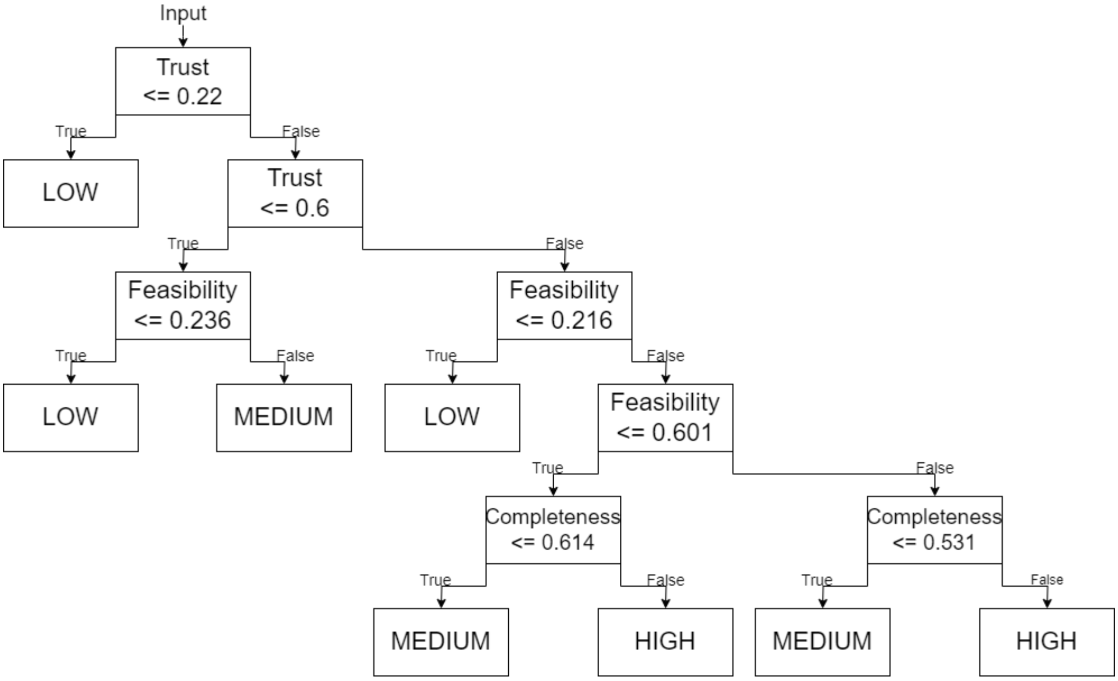


Figure 16. Example of a Decision Tree trained to predict Satisfaction to be low, medium or high. Simplified for displaying purposes from a Decision Tree found in Appendix IV

On the surface, the argumentation displayed in Figure 16 is relatively easy to understand. Initially, the model decides if the explanation given even solves the problem; secondly, if that solution is doable; thirdly, if the explanation is missing any details. While this approach generally proves effective in most cases (achieving an accuracy of 0.79), it overlooks other metrics that might be significant in specific situations.

5.3.3 Feature Importance With Metric-specific Data

To evaluate whether metrics other than Feasibility and Trust are relevant in their specific circumstances, subsets of data corresponding to counterfactuals that targeted these metrics were created. The selection of questions for these subsets was based on those initially intended to target specific metrics and those with metric means significantly deviating from the distribution.

Table 14. Models' accuracies and feature importance when trained on subsets of questions targeting specific metrics.

Metric	Question ID-s	Accuracy	Target metric importance	Feasibility trust importance	Target importance no feasibility or trust	accuracy without trust and feasibility
Consistency	17,19,21,23, 18,20,22,30	0.78	0.061	0.035 0.722	0.621	0.74
Complete.	5,13,14,25, 26,28,20,22, 24,27,30	0.79	0.089	0.162 0.7	0.676	0.67
Fairness	7,16,32,33, 14,31	0.84	0.0	0.639 0.179	0.128	0.84
Complexity	5,28,6,7,8,34	0.67	0.03	0.309 0.566	0.006	0.64

Table 14 shows the questions used to train the Decision Tree models, their performance, and how the targeted metric importance compared to how important Feasibility and Trust were. Furthermore, the targeted metric importances were also tested with models trained without Trust or Feasibility.

With Trust and Feasibility in the training data, the separation of questions did not have an effect, as Feasibility and Trust still held the most importance. However, when removing them from the training process, Consistency and Completeness took the highest importance in their respective models. Fairness and Complexity still remained largely irrelevant though. This is likely due to Consistency and Completeness having the highest remaining correlation with Satisfaction, as seen in Figure 11.

6 Discussion

The analysis in Chapter 5 points towards Feasibility, Trust, Consistency and Completeness being relevant towards predicting overall Satisfaction. Additionally, people seemed to overall agree on how the metrics were exhibited in counterfactual explanations, indicated by the relatively low standard deviations among responses (Table 8). However, the metrics are also heavily correlated amongst themselves (Table 11 and differ greatly in their degree of importance towards predicting Satisfaction. Factor analysis pointed towards the possibility that the variability of the 7 chosen metrics can be explained by 3 underlying factors. However, as the underlying factors correlate with multiple metrics at once, there is no easy method to redefine the factors into new preferences. The most expressive factors having high correlations with all metrics could also indicate a large subgroup of respondents, who simply rated the all metrics similarly for each question.

Looking in depth into how individual answers were rated in Subsection 5.2.2 uncovered the possible root of the high correlations. Namely, the respondents could have let their perception of the general quality of a counterfactual situation affect their ratings of individual metrics. This is a difficult obstacle to overcome, as most people probably do not grade individual aspects of explanations consciously when deciding whether to follow them. There were counterfactuals, where the average scores of metrics did not align with the original expectations in the least, as they were generally rated bad, which in turn lowered all associated metric ratings.

According to machine learning modelling, in most of the cases, satisfaction can successfully be predicted from the metrics measured. The best mean squared error received with a regression model was 0.8, which in a 6-point scale is quite good. Classifying the dataset into 3 classes was also successful, with the accuracy being in the range of 0.75-0.8. This is a significant increase from randomly picking between 3 classes (which would yield an accuracy of 0.33).

The biggest predictor of explanation quality seems to be Feasibility. Even though Trust sometimes held higher weight in models' decisions, individual counterfactual situation analysis in Subsection 5.2.2 exposed, that in situations of potentially high Trust and low Feasibility, the counterfactuals still were rated to have low Trust because of the low Feasibility causing low Satisfaction. Even when separating questions targeting certain metrics as in Table 14, Feasibility was still the most used predictor.

The rest of the metrics take up a lesser role. Completeness and Consistency were sometimes used to separate medium-quality explanations from high quality, like in Figure 16. Fairness remains largely unused for predicting possibly due to its subjectivity, having the highest standard deviation of all metrics (Table 8). Complexity and Understandability played no large role. This could be explained by the low representation of high-Complexity and low-Understandability counterfactuals in the dataset. Other metrics can be successfully used to predict Satisfaction when removing Feasibility and Trust as illustrated in Table 14. However, this may be due to the high correlations that the leftover metrics had with the removed metrics perpetuating the effect Feasibility and Trust had on Satisfaction. The measured metrics clearly are not independent and their effects cannot be definitively separated.

In terms of developing **future counterfactual algorithms**, the clauses added to the search function to make them more useful may be more complicated and layered than the metrics proposed in the questionnaire. As humans' sense of explanation qualities varies and metrics' definitions overlap, there is no easy way to redefine underlying "human-preferred" factors for them to be integrated into code.

If it was necessary to pick a direction, though, the main metric worth integrating into future algorithms is Feasibility, which seems to be a necessity for other metrics even to have an effect. It is easy to reason why it is so: if an explanation is not possible to follow is not worth considering. This is already one of the main directions of counterfactual algorithm development, as there are already multiple algorithms integrating this bias listed by Karimi et al. [17]. Trust is also important, but as stated in Subsection 2.3.4, it has more to do with the machine-learning model accuracy and data. After Feasibility is guaranteed, Completeness and Consistency seem to be the next important steps.

In more practical values, this thesis provides a new dataset where explanatory virtue scores are connected to actual natural-language explanations and the overall quality of the explanation.

The main **limitation** of this thesis is its dependence on data produced by humans not verified by the author himself. The quality of the data used depends on Prolific's [35] sign-up policy and whether faulty data was detected in the data cleaning process. If some respondents misunderstood the metric definitions given in Table 2, the data used will not be ideal for conclusions.

Another limitation is the breadth of domains used in the questionnaire. As the number of questions respondents can answer is limited due to the respondents getting tired, the metric values extracted are only representative of certain areas of life. To build counterfactual search algorithms with the capability of generating useful counterfactual explanations in all situations, domain-specific preferences might need to be researched.

In terms of **future research**, how to integrate actionable feature detection into counterfactual algorithms seems to be a necessary direction. Feasibility in terms of infeasible changes to continuous features has largely been solved by algorithms such as FACE [2] and CARE [22]. However, detecting actionable features has not been automated. As the 3 worst-rated counterfactuals in the questionnaire all changed non-actionable features, this is likely important to users. A possible solution would be to attach a language model to the counterfactual algorithm, which can use feature names to detect features that should not be changed.

Furthermore, as Feasibility seems to be most important, studies with similar questionnaires and metrics, where all counterfactual scenarios are feasible could be run. This eliminates the bias that distracts respondents from impartially evaluating the other metrics. As seen from the trained Decision Tree models, other metrics were only used for predicting Satisfaction once it was sure that the explanation was feasible. From the new questionnaire the next important feature could be discovered and similar modular approach as CARE [22] could be taken to gradually improve counterfactual algorithms to account for all human biases.

7 Conclusion

This thesis set out to discover what humans consider important in counterfactual explanations. To achieve this, a subgroup of criterions (metrics) encompassing human preferences was formed based on previous studies. A questionnaire consisting of factual scenarios and corresponding counterfactual situations was constructed, where respondents were asked to evaluate the counterfactuals based on the aforementioned metrics.

Data analysis was carried out on the responses of the respondents, focusing on two aspects: relations between the metrics and the possibility of predicting overall Satisfaction with the explanation from the metrics. Firstly, it was determined, that the metrics are all tightly interconnected, with Feasibility possibly being the most important explanation quality. Secondly, it was found that predicting overall Satisfaction with the researched metrics can be done relatively accurately.

References

[1] R. K. Mothilal, A. Sharma, and C. Tan, Explaining machine learning classifiers through diverse counterfactual explanations, in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, ser. FAT* '20, ACM, Jan. 2020. DOI: 10.1145/3351095.3372850. (11.5.2024).

[2] R. Poyiadzi, K. Sokol, R. Santos-Rodriguez, T. De Bie, and P. Flach, Face: Feasible and actionable counterfactual explanations, in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '20, ACM, Feb. 2020. DOI: 10.1145/3375627.3375850. (10.4.2024).

[3] R. M. J. Byrne, Counterfactuals in explainable artificial intelligence (xai): Evidence from human reasoning, pp. 6276–6282, Jul. 2019. DOI: 10.24963/ijcai.2019/876. (11.5.2024).

[4] OpenAI, ChatGPT(3.5), https://chat.openai.com, 2022.

[5] Microsoft, Microsoft copilot (copilot free), https://copilot.microsoft.com, 2024.

[6] B. Stratton, Ai industry growth statistics: Exploring key metrics driving growth of ai. [Online]. Available: https://bluetree.digital/ai-industry-growth-metrics/ (22.4.2024).

[7] N. Gillespie, S. Lockey, C. Curtis, J. Pool, and A. Akbari, Trust in artificial intelligence: A global study, *The University of Queensland and KPMG Australia*, 2023. DOI: 10.14264/00d3c94. (12.5.2024).

[8] J. Porter, Chatgpt continues to be one of the fastest-growing services ever, 2023. [Online]. Available: https://www.theverge.com/2023/11/6/23948386/chatgpt-active-user-count-openai-developer-conference (12.5.2024).

[9] C. Piers, Even chatgpt says chatgpt is racially biased, Feb. 2024. [Online]. Available: https://www.scientificamerican.com/article/even-chatgpt-says-chatgpt-is-racially-biased/ (22.4.2024).

[10] S. Anwarul and S. Dahiya, A comprehensive review on face recognition methods and factors affecting facial recognition accuracy, pp. 495–514, 2020. DOI: 10.1007/978-3-030-29407-6_36. (22.4.2024).

[11] M. T. Ribeiro, S. Singh, and C. Guestrin, "why should i trust you?" explaining the predictions of any classifier, 2016. arXiv: 1602.04938 [cs.LG]. (11.5.2024).

[12] S. Lundberg and S.-I. Lee, A unified approach to interpreting model predictions, 2017. arXiv: 1705.07874 [cs.AI]. (11.5.2024).

[13] H. Blockeel, L. Devos, B. Frénay, G. Nanfack, and S. Nijssen, Decision trees: From efficient prediction to responsible ai, *Frontiers in Artificial Intelligence*, vol. 6, 2023. DOI: 10.3389/frai.2023.1124553.

[14] R. K. Mothilal, A. Sharma, and C. Tan, Diverse counterfactual explanations (dice) for ml. [Online]. Available: https://interpret.ml/DiCE/ (22.4.2024).

[15] M. Domnich and R. Vicente, Enhancing counterfactual explanation search with diffusion distance and directional coherence, 2024. arXiv: 2404.12810 [cs.LG]. (11.5.2024).

[16] S. Wachter, B. Mittelstadt, and C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the gdpr, 2018. arXiv: 1711.00399 [cs.AI]. (5.2.2024).

[17] A.-H. Karimi, G. Barthe, B. Schölkopf, and I. Valera, A survey of algorithmic recourse: Contrastive explanations and consequential recommendations, *Association for Computing Machinery Comput. Surv.*, vol. 55, no. 5, Dec. 2022. DOI: 10.1145/3527848. (9.4.2024).

[18] B. Ustun, A. Spangher, and Y. Liu, Actionable recourse in linear classification, in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ser. FAT* '19, ACM, Jan. 2019. DOI: 10.1145/3287560.3287566. (6.2.2024).

[19] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, *Artificial Intelligence*, vol. 267, pp. 1–38, 2019. DOI: https://doi.org/10.1016/j.artint.2018.07.007. (11.5.2024).

[20] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, Measures for explainable ai: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-ai performance, *Frontiers in Computer Science*, vol. 5, 2023. DOI: 10.3389/fcomp.2023.1096257. (15.4.2024).

[21] J. C. Zemla, S. Sloman, C. Bechlivanidis, and D. A. Lagnado, Evaluating everyday explanations, *Psychonomic Bulletin & Review*, vol. 24, no. 5, pp. 1488–1500, Oct. 2017. DOI: 10.3758/s13423-017-1258-z. (12.4.2024).

[22] P. Rasouli and I. C. Yu, Care: Coherent actionable recourse based on sound counterfactual explanations, 2021. arXiv: 2108.08197 [cs.LG]. (12.4.2024).

[23] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, and S. Goel, Racial disparities in automated speech recognition, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 117, no. 14, pp. 7684–7689, Apr. 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7149386/ (15.4.2024).

[24] R. Shams, D. Zowghi, and M. Bano, Ai and the quest for diversity and inclusion: A systematic literature review, *AI and Ethics*, pp. 1–28, Nov. 2023. DOI: 10.1007/s43681-023-00362-w. (9.4.2024).

[25] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, 2018. DOI: https://doi.org/10.48550/arXiv.1706.07269. arXiv: 1706.07269 [cs.AI]. (11.5.2024).

[26] T. Lombrozo, Simplicity and probability in causal explanation, *Cognitive Psychology*, vol. 55, no. 3, pp. 232–257, 2007. DOI: https://doi.org/10.1016/j.cogpsych.2006.09.006. (11.5.2024).

[27] S. Verma, V. Boonsanong, M. Hoang, K. E. Hines, J. P. Dickerson, and C. Shah, Counterfactual explanations and algorithmic recourses for machine learning: A review, 2022. arXiv: 2010.10596 [cs.LG]. (11.5.2024).

[28] M. Pawelczyk, K. Broelemann, and G. Kasneci, On counterfactual explanations under predictive multiplicity, 2020. arXiv: 2006.13132 [cs.LG]. (16.4.2024).

[29] S. Sharma, J. Henderson, and J. Ghosh, Certifai: A common framework to provide explanations and analyse the fairness and robustness of black-box models, in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '20, ACM, Feb. 2020. DOI: 10.1145/3375627.3375812. (11.5.2024).

[30] J. Välja, Assessing the quality of counterfactual explanations with large language models, University of Tartu, Institute of Computer Science, Bachelor's thesis, 2024.

[31] B. Becker and R. Kohavi, Adult, UCI Machine Learning Repository, 1996. [Online]. Available: https://doi.org/10.24432/C5XW20 (11.5.2024).

[32] Pima indians diabetes database. [Online]. Available: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database (23.4.2024).

[33] A. Vaish, T. Grossmann, and A. Woodward, Not all emotions are created equal: The negativity bias in social-emotional development. en, *Psychological Bulletin*, vol. 134, no. 3, pp. 383–403, May 2008. DOI: 10.1037/0033-2909.134.3.383. (11.5.2024).

[34] Limesurvey, https://survey.ut.ee/.

[35] Prolific, https://www.prolific.com/.

[36] K. Pearson, Liii. on lines and planes of closest fit to systems of points in space, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901. DOI: 10.1080/14786440109462720.

[37] H. Hotelling, Analysis of a complex of statistical variables into principal components, *J. Educ. Psychol.*, vol. 24, no. 6, pp. 417–441, 1933. DOI: 10.1037/h0071325. (11.5.2024).

[38] L. van der Maaten and G. Hinton, Viualizing data using t-sne, *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, Nov. 2008. [Online]. Available: https://www.researchgate.net/publication/228339739_Viualizing_data_using_t-SNE (11.5.2024).

[39] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, Portland, Oregon: AAAI Press, 1996, pp. 226–231. [Online]. Available: https://dl.acm.org/doi/10.5555/3001460.3001507 (11.5.2024).

[40] J. A. Hartigan, Direct clustering of a data matrix, *Journal of the American Statistical Association*, vol. 67, no. 337, pp. 123–129, 1972. [Online]. Available: http://www.jstor.org/stable/2284710 (23.4.2024).

[41] S. Cinaroglu, Comparison of performance of decision tree algorithms and random forest: An application on oecd countries health expenditures, *International Journal of Computer Applications*, vol. 138, pp. 37–41, Mar. 2016. DOI: 10.5120/ijca2016908704. (11.5.2024).

[42] M. V. Datla, Bench marking of classification algorithms: Decision trees and random forests - a case study using r, in *2015 International Conference on Trends in Automation, Communications and Computing Technology (I-TACT-15)*, 2015, pp. 1–7. DOI: 10.1109/ITACT.2015.7492647. (11.5.2024).

[43] G. Batista, R. Prati, and M.-C. Monard, A study of the behavior of several methods for balancing machine learning training data, *SIGKDD Explorations*, vol. 6, pp. 20–29, Jun. 2004. DOI: 10.1145/1007730.1007735. (11.5.2024).

[44] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, Smote: Synthetic minority over-sampling technique, *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002. DOI: 10.1613/jair.953. (11.5.2024).

[45] I. Tomek, Two modifications of cnn, *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-6, no. 11, pp. 769–772, 1976. DOI: 10.1109/TSMC.1976.4309452. (11.5.2024).

[46] B. Raad and B. Mlacic, Big five factor model, theory and structure, in Dec. 2015, pp. 559–566. DOI: 10.1016/B978-0-08-097086-8.25066-6. (11.5.2024).

Appendix

I. Questionnaire

ID 5

Imagine you are in this scenario:

"You are a 31-year-old divorced woman. You have a high-school education and you work 20 hours per week."

Current outcome: You are earning **less than the average salary.**

Useful context: the standard full-time workload is 40 hours per week.

To **earn more than the average salary**, you would need to make the following changes:"

- Increase your **education level** from **high-school** to **Bachelor's degree**.

The rest of the values will remain constant.

ID 6

Imagine you are in this scenario:

"You are a 31-year-old divorced woman. You have a high-school education and you work 20 hours per week."

Current outcome: You are earning **less than the average salary.**

Useful context: the standard full-time workload is 40 hours per week.

To **earn more than the average salary**, you would need to make the following changes:

- Increase your **education level** from **high-school** to **Bachelor's degree**.
- And, increase your **hours worked per week** from **20** to **30**.

The rest of the values will remain constant.

ID 7

Imagine you are in this scenario:

"You are a 31-year-old divorced woman. You have a high-school education and you work 20 hours per week."

Current outcome: You are earning **less than the average salary.**

Useful context: the standard full-time workload is 40 hours per week.

To **earn more than the average salary**, you would need to make the following changes:

- Increase your **education level** from **high-school** to **Bachelor's degree**.
- And, increase your **hours worked per week** from **20** to **30**.
- And, change your **relationship status** from **divorced** to **married**.

The rest of the values will remain constant.

ID 8

Imagine you are in this scenario:

"You are a 31-year-old divorced woman. You have a high-school education and you work 20 hours per week as a junior accountant. You live in France in a two-room apartment. You have a pet cat and exercise regularly."

Current outcome: You are earning **less than the average salary.**

Useful context: the standard full-time workload is 40 hours per week.

To **earn more than the average salary**, you would need to make the following changes:

- Increase your **education level** from **high-school** to **Bachelor's degree**.
- And, increase your **hours worked per week** from **20** to **30**.
- And, change your **relationship status** from **divorced** to **married**.
- And, change your **exercising habits** from **regular** to **occasional**.
- And, change your **occupation** from **junior accountant** to **senior accountant**.

The rest of the values will remain constant.

ID 9

Imagine you are in this scenario:

"You are a 23-year-old single woman with a child. You have a college education and you work 20 hours per week as a barista. You have no passive income."

Current outcome: You are earning **less than the average salary.**

Useful context: the standard full-time workload is 40 hours per week. Passive income can consist of sources of income such as owning stocks or renting out living spaces.

To **earn more than the average salary**, you would need to make the following changes:

- Increase your **working hours per week** from **20** to **95**.

The rest of the values will remain constant.

ID 10

Imagine you are in this scenario:

"You are a 23-year-old single woman with a child. You have a college education and you work 20 hours per week as a barista. You have no passive income."

Current outcome: You are earning **less than the average salary.**

Useful context: the standard full-time workload is 40 hours per week. Passive income can consist of sources of income such as owning stocks or renting out living spaces.

To **earn more than the average salary**, you would need to make the following changes:

- Increase your **working hours per week** from **20** to **30**.

- And, increase your **passive income** from **0$** to **1500$** per month.

The rest of the values will remain constant.

ID 11

Imagine you are in this scenario:

"You are a 23-year-old single woman with a child. You have a college education and you work 20 hours per week as a barista. You have a passive income of 1500$ per month"

Current outcome: You are earning **less than the average salary.**

Useful context: the standard full-time workload is 40 hours per week. Passive income can consist of sources of income such as owning stocks or renting out living spaces.

To **earn more than the average salary**, you would need to make the following changes:

- Increase your **working hours per week** from **20** to **30**.
- Increase your **passive income** from **1500$** to **3000$** per month.

The rest of the values will remain constant.

ID 12

Imagine you are in this scenario:

"You are a 23-year-old single woman with a child. You have a college education and you work 20 hours per week as a barista. You have no passive income."

Current outcome: You are earning **less than the average salary.**

Useful context: the standard full-time workload is 40 hours per week. Passive income can consist of sources of income such as owning stocks or renting out living spaces.

To **earn more than the average salary**, you would need to make the following changes:

- Increase your **working hours per week** from **20** to **60**.

The rest of the values will remain constant.

ID 13

Imagine you are in this scenario:

"You are a 23-year-old single woman with a child. You have a high school education and you work 20 hours per week as a barista. You have no passive income."

Current outcome: You are earning **less than the average salary.**

Useful context: the standard full-time workload is 40 hours per week. Passive income can consist of sources of income such as owning stocks or renting out living spaces.

To **earn more than the average salary**, you would need to make the following changes:

- Increase your **education level** from **high school** to **PhD**.

The rest of the values will remain constant.

ID 14

Imagine you are in this scenario:

"You are a 33-year-old single woman with a child. You have a PhD and you work 20 hours per week as a barista. You have no passive income."

Current outcome: You are earning **less than the average salary.**

Useful context: the standard full-time workload is 40 hours per week. Passive income can consist of sources of income such as owning stocks or renting out living spaces.

To **earn more than the average salary**, you would need to make the following changes:

- Decrease your **education level** from **PhD** to **Master's**.

The rest of the values will remain constant.

ID 15

Imagine you are in this scenario:

"You are a 23-year-old single woman with a child. You have a college education and you work 20 hours per week as a barista. You have no passive income."

Current outcome: You are earning **less than the average salary.**

Useful context: the standard full-time workload is 40 hours per week. Passive income can consist of sources of income such as owning stocks or renting out living spaces.

To **earn more than the average salary**, you would need to make the following changes:

- Change your **occupation** from **barista** to **hairdresser**.

The rest of the values will remain constant.

ID 16

Imagine you are in this scenario:

"You are a 23-year-old single woman with a child. You have a college education and you work 20 hours per week as a barista. You have no passive income and you were born in Germany."

Current outcome: You are earning **less than the average salary.**

Useful context: the standard full-time workload is 40 hours per week. Passive income can consist of sources of income such as owning stocks or renting out living spaces.

To **earn more than the average salary**, you would need to make the following changes:

- Change the **country** you were born in from **Germany** to **Sweden**.

The rest of the values will remain constant.

ID 17

Imagine you are in this scenario:

"You are a 31-year-old person, who has had one pregnancy, and who has the following medical readings: glucose level: 173 mg/ dL, diastolic blood pressure: 82 mm Hg, skin thickness: 48 mm, Insulin: 160 μIU/ml, body mass index (BMI): 32.8."

Current outcome: You have an **increased risk of diabetes**.

Normal glucose levels are 70-130, 140-200 is prediabetes, 200+ is diabetes. Healthy BMI is 18.5 - 25, 25-30 is considered overweight, 30+ is considered obese. Normal diastolic blood pressure is roughly below 80 mm Hg. Normal insulin is between 16 and 166 μIU/ml.

To **not be at increased risk of diabetes** you would need to make the following changes:

- Decrease your **glucose level** from **173** to **130**.
- And, increase your **insulin level** from **160** to **181**.

The rest of the values will remain constant.

ID 18

Imagine you are in this scenario:

"You are a 31-year-old person, who has had one pregnancy, and who has the following medical readings: glucose level: 173 mg/dL, diastolic blood pressure: 82 mm Hg, skin thickness: 48 mm, Insulin: 160 μIU/ml, body mass index (BMI): 32.8."

Current outcome: You have an **increased risk of diabetes**.

Normal glucose levels are 70-130, 140-200 is prediabetes, 200+ is diabetes. Healthy BMI is 18.5 - 25, 25-30 is considered overweight, 30+ is considered obese. Normal diastolic blood pressure is roughly below 80 mm Hg. Normal insulin is between 16 and 166 μIU/ml.

To **not be at increased risk of diabetes** you would need to make the following changes:

- Decrease your **glucose level** from **173** to **135**.
- And, decrease your **insulin level** from **160** to **142**.

The rest of the values will remain constant.

ID 19

Imagine you are in this scenario:

"You are a 27-year-old patient, who took a medium-sized dose of a drug for treating the flu. You work as a bartender and smoke occasionally."

Current outcome: The **duration of your flu** is **not reduced**.

To **reduce the duration of your flu**, you would need to make the following changes:

- Administer a **high dose** of the **drug** instead of a **medium dose**.
- And, adjust your **smoking habits** from **occasional** to **regular**.

The rest of the values will remain constant.

ID 20

Imagine you are in this scenario:

"You are a 27-year-old patient, who took a medium-sized dose of a drug for treating the flu. You work as a bartender and smoke occasionally."

Current outcome: The **duration of your flu** is **not reduced**.

To **reduce the duration of your flu**, you would need to make the following changes:

- Administer a **high dose** of the **drug** instead of a **medium dose**.
- And, adjust your **smoking habits** from **occasional** to **rare**.

The rest of the values will remain constant.

ID 21

Imagine you are in this scenario:

"You are a 19-year-old student who studies for 25 hours per week and has an average grade of B. You are proficient in English and are seeking to apply to a well-regarded university."

Current outcome: You were **rejected from the university**.

To **be accepted into the university**, you would need to make the following changes:

- Increase your **hours studied per week** from **25** to **31**.
- And, lower your **average grade** from **B** to **C**.

The rest of the values will remain constant.

ID 22

Imagine you are in this scenario:

"You are a 19-year-old student who studies for 25 hours per week and has an average grade of B. You are proficient in English and are seeking to apply to a well-regarded university."

Current outcome: You were **rejected from the university**.

To **be accepted into the university**, you would need to make the following changes:

- Increase your **hours studied per week** from **25** to **31**.
- And, improve your **average grade** from **B** to **A**.

The rest of the values will remain constant.

ID 23

Imagine you are in this scenario:

"You are a 19-year-old student who studies for 25 hours per week and has an average grade of B. You are proficient in English and are seeking to apply to a well-regarded university."

Current outcome: You were **rejected from the university**.

To **be accepted into the university**, you would need to make the following changes:

- Decrease your **hours studied per week** from **25** to **22**.
- And, improve your **average grade** from **B** to **A**.

The rest of the values will remain constant.

ID 24

Imagine you are in this scenario:

"You are a 38-year-old woman, who is working as a cashier and has a Bachelor's degree in Communications. You work for 25 hours per week. You were born in Germany and your relationship status is married."

Current outcome: You are earning **less than the average salary**.

Useful context: the standard full-time workload is 40 hours per week.

To **earn more than the average salary**, you would need to make the following changes:

- Increase your **hours worked per week** from **25** to **30**.
- And, change your **job position** from **cashier** to **store manager**.

The rest of the values will remain constant.

ID 25

Imagine you are in this scenario:

"You are a 38-year-old person, who is working as a cashier and has a Bachelor's degree in Communications. You were born in Germany and your relationship status is married."

Current outcome: You are earning **less than the average salary**.

To **earn more than the average salary**, you would need to make the following changes:

- Obtain a **Master's degree in Communications** in addition to your **Bachelor's degree**.

The rest of the values will remain constant.

ID 26

Imagine you are in this scenario:

"You are a 38-year-old person, who is working as a junior researcher and has a Bachelor's degree in Biology. You were born in Germany and your relationship status is married."

Current outcome: You are earning **less than the average salary**.

To **earn more than the average salary**, you would need to make the following changes:

- Change your **occupation** from **junior researcher** to **lawyer**.

The rest of the values will remain constant.

ID 27

Imagine you are in this scenario:

"You are a 31-year-old person who works at a law firm in a junior position. You have an education level of Master's. You work for 65 hours per week and work overtime on weekends. The firm offers a stress-management course in which you are not participating at the moment."

Current outcome: You are **at high risk of experiencing burnout**.

To **not have a high risk of experiencing burnout**, you would need to make the following changes:

- Take part in a **stress-management course**.

The rest of the values will remain constant.

ID 28

Imagine you are in this scenario:

"You are a 38-year-old person who lives in Sweden. You work from home as a graphic designer. You do not exercise regularly. Your daily routine involves long hours of sitting. Additionally, your diet is made up of unhealthy food choices, leading to calorie intake that significantly exceeds burning. You also experience food cravings that contribute to irregular and excessive eating habits."

Current outcome: You have an **increased risk of obesity-related health problems**.

To **not be at high risk of obesity-related health problems**, you would need to make the following changes:

- Start **working out once a month**.

The rest of the values will remain constant.

ID 29

Imagine you are in this scenario:

"You are a 40-year-old software developer with excellent technical skills. You have a good reputation for technical problem-solving, but you have not had any project where you took a leadership role or worked on cross-team collaboration. You have not actively participated in industry networking events and have not taken any leadership training opportunities."

Current outcome: You have **not been promoted to a managerial position**.

To **be promoted to a managerial position**, you would need to make the following changes:

- Take an **intro course to project management**.

The rest of the values will remain constant.

ID 30

Imagine you are in this scenario:

"You are a 40-year-old software developer with excellent technical skills. You have a good reputation for technical problem-solving, but you have not had any projects where you took a leadership role or worked on cross-team collaboration. You have not actively participated in industry networking events and have not taken any leadership training opportunities."

Current outcome: You have **not been promoted to a managerial position**.

To **be promoted to a managerial position**, you would need to make the following changes:

- Take an **intro course to project management**.
- Gain **leadership experience** by **leading a team or project**.
- Attend **industry networking events**.

The rest of the values will remain constant.

ID 31

Imagine you are in this scenario:

"You are a 25-year-old man, who is working as a junior developer and has finished high school. You were born in the UK and you are divorced. You work 20 hours per week."

Current outcome: You **did not get the software developer job you applied for**.

To **get the software developer job you applied for**, you would need to make the following changes:

- Increase your **education level** from **high school** to **Bachelor's**.
- Change your current **job position** from **junior developer** to **senior developer**.

The rest of the values will remain constant.

ID 32

Imagine you are in this scenario:

"You are a 22-year-old man, who is working as a junior developer and has finished high school. You were born in the UK and you are divorced. You work 20 hours per week."

Current outcome: You **did not get the software developer job you applied for**.

To **get the software developer job you applied for**, you would need to make the following changes:

- Change your **gender** from **male** to **female**.
- And, increase your **working hours per week** from **20** to **40**.

The rest of the values will remain constant.

ID 33

Imagine you are in this scenario:

"You are a 22-year-old single man, who is working as a junior developer and has finished high school. You were born in the UK and you work for 20 hours per week."

Current outcome: You **did not get the software developer job you applied for**.

To **get the software developer job you applied for**, you would need to make the following changes:

- Change your **age** from **22** to **25**.
- And, change your **relationship status** from **single** to **married**.

The rest of the values will remain constant.

ID 34

Imagine you are in this scenario:

"You are a 41-year-old married man, who works as a CFO in a medium-sized bank and has finished high school. You were born in the UK and you work for 35 hours per week. Your income is $5100 per month and your savings account contains $140,000. You have a credit history FICO score of 685. You ask for a loan term of 360 months."

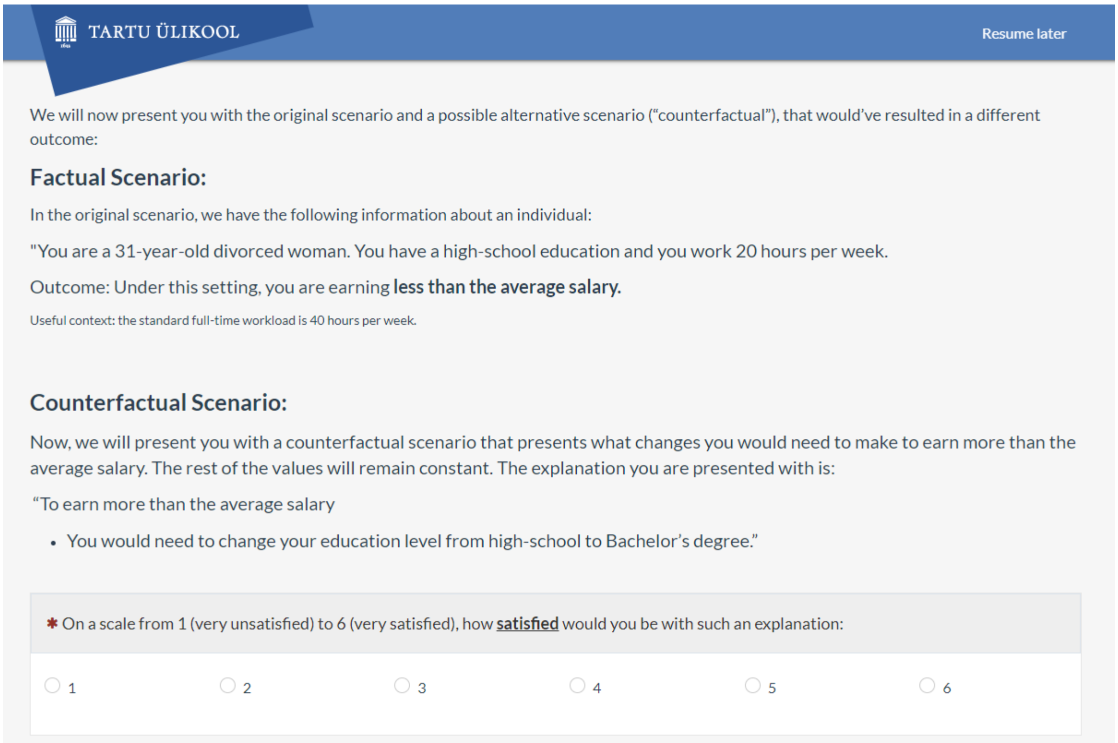
Current outcome: You **did not get a loan to start a new company**.

Useful context: FICO credit history score 300-579 is considered a poor score, 579-669 - fair, 669-739 - good score, 739-799 - very good and above 799 - exceptional.

To **get the loan to start a new company**, you would need to make the following changes:

- Change your **income** from **$5100** to **$5150** per month.
- And, change your **savings balance** from **$140,000** to **$144,000**.
- And, increase your **hours worked per week** from **35** to **37**.
- And, increase your **age** from **41** to **43**.
- And, improve your **credit score** from **685** to **690**.
- And, change the **loan term amount** from **360 months** to **370 months**.

The rest of the values will remain constant.



We will now present you with the original scenario and a possible alternative scenario ("counterfactual"), that would've resulted in a different outcome:

Factual Scenario:

In the original scenario, we have the following information about an individual:

"You are a 31-year-old divorced woman. You have a high-school education and you work 20 hours per week.

Outcome: Under this setting, you are earning **less than the average salary.**

Useful context: the standard full-time workload is 40 hours per week.

Counterfactual Scenario:

Now, we will present you with a counterfactual scenario that presents what changes you would need to make to earn more than the average salary. The rest of the values will remain constant. The explanation you are presented with is:

"To earn more than the average salary

- You would need to change your education level from high-school to Bachelor's degree."

* On a scale from 1 (very unsatisfied) to 6 (very satisfied), how **satisfied** would you be with such an explanation:

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6

Figure I.1. Example of how questions were presented in the pilot study.

Examples of counterfactuals provided to respondents to help their grading

1. Feasibility

   - Good: "If you decrease your mobile phone use from 6 hours a day to 5.5 hours a day, you'll likely experience less symptoms of anxiety."
   - Bad: Changes an attribute that cannot be changed. "If you decrease your age from 42 to 26, you will reduce your risk of a heart attack."

2. Consistency

   - Good: "You drove to a concert at an average of 98 km/h and got pulled over by the police 2 times. You left for the concert at 17:30. As a result, you were late for the concert. If you had left for the concert at 17:15, you would have made it to the concert on time."
   - Bad: The suggested changes are contradictory. "You drove to a concert at an average of 98 km/h and got pulled over by the police 2 times. As a result, you were late for the concert. If you had driven at an average of 104 km/h and got pulled over by the police 0 times, you would have made it to the concert on time."

3. Completeness

- Scenario: "You are a 20 year old woman, who has an average grade of 81%. You applied to university without the required motivation letter and meeting the average grade threshold of 90%. You were not accepted into the university."
- Good: "If you had an average grade of 91% and had written a motivation letter, you would have been accepted into the university."
- Bad: The suggested changes skip important steps for achieving the desired outcome. "If you had written a motivation letter, you would have been accepted into the university."

4. I believe that the suggested changes would bring about the desired outcome.

   - Good: "If you had left home 15 minutes earlier, you would have caught the earlier train, as per the schedule."
   - Bad: The suggested change is unlikely to bring about the desired change. "If you called the train conductor to wait an extra 15 minutes, you wouldn't have missed the train."

5. Understandability

   - Good: "If the patient had arrived ten minutes earlier, immediate treatment could have prevented the cardiac arrest."
   - Bad: The explanation uses overly specific and confusing phrases. "The patient's arrival time juxtaposed with the chronological treatment window delineates an alternate outcome scenario."

6. Fairness

   - Good: "Applicants with a minimum of 3 years of experience have a higher chance of getting hired for this role."
   - Bad: The explanation suggests changes that would be commonly viewed as unfair or illegal. "People who are over 35 are not usually accepted for this position."

7. Complexity

   - Good: "Reducing your debt by 10% could improve your credit rating to meet our loan approval criteria."
   - Too complex: The explanation is unnecessarily long and complex."To improve your credit score to a level that matches our loan approval benchmarks, you should decrease your debt by 4%, increase your monthly savings by 41$, increase your monthly income by 23$, increase your weekly work hours from 40 to 42, work at your current company for 1 more year and receive a letter of recommendation from your current employer.

II. Code

Code used to produce the results in this thesis found can be found in github:
https://github.com/RasmusVeski/Measuring_Human_Preferences_in_Counterfactual_Explanations/tree/main.

III. Questionnaire Analysis

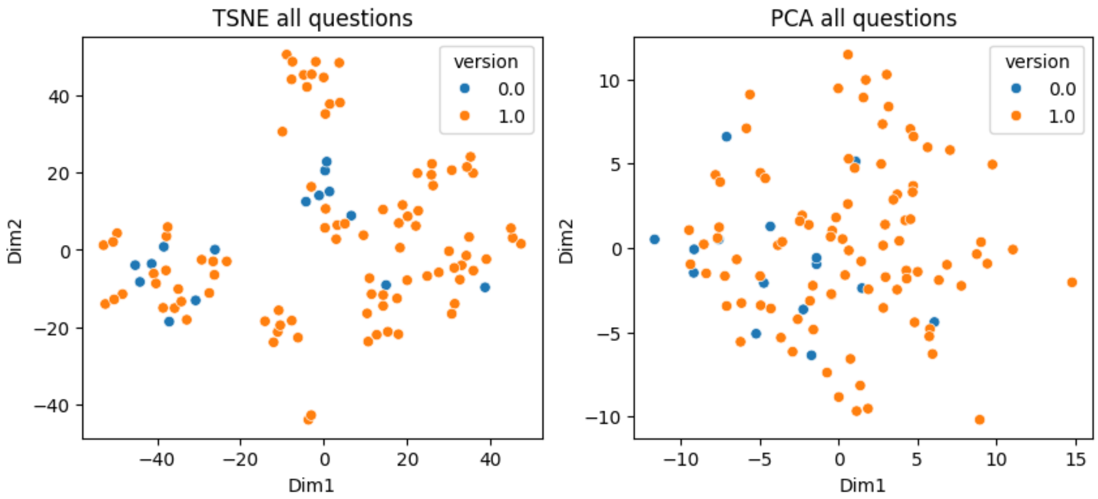


Figure III.1. In order to detect if the respondents to the pilot study rated differently than the respondents in the main study, they were compared. Plotted respondents answers reduced to 2 dimensions by t-SNE and PCA. Blue dots are respondents from pilot study, orange main study

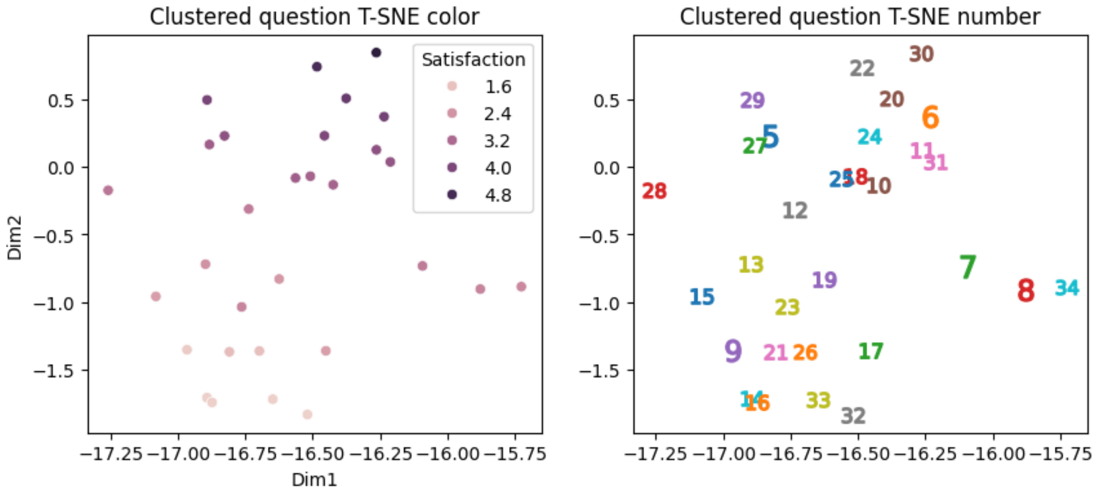


Figure III.2. Average question metric values reduced to 2 dimensions (t-SNE perplexity 25).In the left image the questions are colored by average Satisfaction, in the right the corresponding question id-s

Table III.1. Counterfactual mean values by metric structured "counterfactualID : mean", ordered separately by column by mean values

Satisfaction	Feasibility	Consist.	Complet.	Trust	Understand.	Fairness	Complexity
30 : 5.19	30 : 5.07	30 : 5.49	30 : 5.32	30 : 5.33	22 : 5.53	30 : 5.4	34 : 0.9
22 : 4.76	29 : 5.04	22 : 5.32	22 : 4.87	22 : 4.89	30 : 5.49	22 : 5.29	8 : 0.58
20 : 4.33	22 : 4.95	29 : 5.02	20 : 4.53	6 : 4.52	29 : 5.49	29 : 5.14	30 : 0.14
6 : 4.12	20 : 4.55	20 : 4.82	6 : 4.33	20 : 4.46	5 : 5.4	20 : 5.01	7 : 0.12
29 : 4.04	27 : 4.52	6 : 4.82	11 : 4.14	24 : 4.27	27 : 5.33	28 : 4.73	6 : -0.03
24 : 3.97	28 : 4.44	5 : 4.73	24 : 4.04	31 : 4.13	6 : 5.31	6 : 4.7	31 : -0.04
5 : 3.75	5 : 4.18	27 : 4.7	31 : 4.02	29 : 4.13	24 : 5.31	24 : 4.67	11 : -0.13
11 : 3.75	6 : 4.12	24 : 4.58	8 : 3.76	11 : 4.11	20 : 5.31	31 : 4.63	20 : -0.14
31 : 3.64	18 : 3.82	11 : 4.55	10 : 3.76	5 : 4.0	11 : 5.21	5 : 4.62	17 : -0.18
27 : 3.45	25 : 3.77	18 : 4.48	34 : 3.73	10 : 3.95	28 : 5.19	11 : 4.6	18 : -0.22
25 : 3.44	24 : 3.71	25 : 4.41	29 : 3.71	18 : 3.81	25 : 5.14	27 : 4.58	25 : -0.23
18 : 3.35	19 : 3.57	10 : 4.34	5 : 3.59	25 : 3.54	31 : 5.09	18 : 4.57	22 : -0.23
10 : 3.33	31 : 3.54	12 : 4.22	27 : 3.57	12 : 3.51	12 : 5.08	25 : 4.44	24 : -0.26
28 : 2.95	11 : 3.42	28 : 4.16	12 : 3.53	27 : 3.48	10 : 5.07	10 : 4.22	10 : -0.27
7 : 2.84	10 : 3.09	31 : 4.15	7 : 3.48	7 : 3.05	18 : 4.93	23 : 4.04	32 : -0.32
12 : 2.82	23 : 2.99	13 : 3.87	25 : 3.45	13 : 3.0	13 : 4.84	19 : 3.86	19 : -0.37
8 : 2.68	15 : 2.95	7 : 3.65	18 : 3.44	8 : 2.74	19 : 4.78	17 : 3.8	33 : -0.44
23 : 2.65	21 : 2.95	15 : 3.32	19 : 3.31	34 : 2.74	7 : 4.76	12 : 3.74	26 : -0.44
34 : 2.57	12 : 2.86	34 : 3.2	23 : 2.84	28 : 2.66	15 : 4.67	15 : 3.74	23 : -0.49
13 : 2.41	7 : 2.82	9 : 3.0	28 : 2.73	19 : 2.25	8 : 4.63	13 : 3.66	5 : -0.54
19 : 2.31	13 : 2.56	8 : 2.91	13 : 2.65	9 : 2.22	23 : 4.58	21 : 3.63	21 : -0.54
15 : 2.22	17 : 2.55	26 : 2.49	21 : 2.35	26 : 2.21	9 : 4.58	34 : 3.62	27 : -0.56
17 : 2.2	8 : 2.51	19 : 2.49	9 : 2.34	15 : 2.19	26 : 4.52	26 : 3.2	12 : -0.57
21 : 1.69	34 : 2.47	23 : 2.29	17 : 2.34	17 : 2.14	34 : 4.33	14 : 3.01	13 : -0.62
26 : 1.64	16 : 1.64	16 : 2.25	15 : 2.29	23 : 2.09	21 : 4.26	8 : 3.0	29 : -0.67
9 : 1.44	26 : 1.62	17 : 2.16	32 : 2.16	21 : 1.62	33 : 4.2	7 : 2.88	16 : -0.73
14 : 1.37	33 : 1.38	33 : 2.08	26 : 2.15	16 : 1.51	32 : 4.14	9 : 2.87	14 : -0.74
33 : 1.37	14 : 1.35	32 : 1.92	33 : 2.01	32 : 1.43	17 : 4.09	33 : 2.32	9 : -0.75
16 : 1.36	9 : 1.34	14 : 1.89	14 : 1.86	14 : 1.42	16 : 4.09	16 : 2.22	15 : -0.89
32 : 1.33	32 : 1.32	21 : 1.7	16 : 1.71	33 : 1.36	14 : 4.03	32 : 1.52	28 : -1.09

Table III.2. Counterfactual standard deviation values by metric structured "counterfactualID : stdv", ordered separately by column by standard deviation value

Satisfaction	Feasibility	Consist.	Complet.	Trust	Understand.	Fairness	Complexity
28 : 1.61	28 : 1.67	9 : 1.8	28 : 1.71	9 : 1.65	16 : 2.03	14 : 2.01	16 : 1.43
27 : 1.59	19 : 1.67	34 : 1.69	34 : 1.69	28 : 1.65	14 : 1.96	9 : 2.01	9 : 1.41
24 : 1.5	27 : 1.64	26 : 1.63	29 : 1.69	27 : 1.64	33 : 1.92	21 : 1.91	26 : 1.34
29 : 1.48	21 : 1.61	16 : 1.63	19 : 1.64	12 : 1.61	32 : 1.85	19 : 1.89	14 : 1.34
10 : 1.48	15 : 1.52	15 : 1.62	27 : 1.62	13 : 1.61	21 : 1.76	26 : 1.89	32 : 1.32
25 : 1.48	24 : 1.48	13 : 1.61	9 : 1.61	34 : 1.56	9 : 1.73	17 : 1.81	13 : 1.3
19 : 1.47	34 : 1.45	28 : 1.59	10 : 1.61	17 : 1.56	17 : 1.62	23 : 1.79	17 : 1.28
31 : 1.45	31 : 1.43	19 : 1.59	8 : 1.57	7 : 1.55	15 : 1.61	12 : 1.72	18 : 1.21
12 : 1.45	12 : 1.42	12 : 1.55	24 : 1.55	29 : 1.54	26 : 1.58	13 : 1.69	33 : 1.18
18 : 1.45	6 : 1.42	8 : 1.53	25 : 1.5	25 : 1.53	34 : 1.56	15 : 1.65	8 : 1.16
23 : 1.41	10 : 1.4	17 : 1.49	5 : 1.5	19 : 1.52	13 : 1.55	27 : 1.65	34 : 1.15
11 : 1.39	13 : 1.36	31 : 1.47	13 : 1.49	18 : 1.5	19 : 1.5	16 : 1.64	25 : 1.09
8 : 1.38	25 : 1.34	25 : 1.45	21 : 1.49	11 : 1.49	23 : 1.42	10 : 1.6	15 : 1.06
34 : 1.37	11 : 1.33	33 : 1.45	12 : 1.46	10 : 1.48	8 : 1.31	33 : 1.6	21 : 1.04
6 : 1.36	18 : 1.32	7 : 1.43	11 : 1.46	5 : 1.44	7 : 1.29	34 : 1.6	5 : 1.03
13 : 1.33	23 : 1.3	10 : 1.43	18 : 1.42	24 : 1.43	12 : 1.25	8 : 1.56	28 : 1.02
5 : 1.31	20 : 1.27	23 : 1.39	32 : 1.41	26 : 1.42	18 : 1.22	7 : 1.54	12 : 1.01
15 : 1.3	8 : 1.26	27 : 1.38	7 : 1.4	8 : 1.37	28 : 1.2	6 : 1.49	27 : 1.0
7 : 1.27	5 : 1.26	32 : 1.38	31 : 1.4	20 : 1.33	20 : 1.18	5 : 1.45	11 : 0.99
20 : 1.27	16 : 1.24	14 : 1.37	15 : 1.35	31 : 1.28	10 : 1.15	28 : 1.43	7 : 0.98
22 : 1.22	17 : 1.23	11 : 1.34	6 : 1.34	6 : 1.27	31 : 1.12	24 : 1.41	22 : 0.97
17 : 1.21	7 : 1.23	18 : 1.33	23 : 1.31	15 : 1.26	25 : 1.05	25 : 1.36	29 : 0.97
21 : 1.09	29 : 1.08	24 : 1.32	17 : 1.28	23 : 1.22	24 : 1.01	18 : 1.33	23 : 0.92
9 : 0.97	22 : 0.97	20 : 1.24	26 : 1.28	22 : 1.15	27 : 1.01	11 : 1.33	10 : 0.92
30 : 0.95	26 : 0.95	5 : 1.23	33 : 1.28	21 : 1.14	11 : 0.96	31 : 1.32	20 : 0.91
33 : 0.93	9 : 0.91	6 : 1.22	22 : 1.22	16 : 0.94	6 : 0.95	20 : 1.26	19 : 0.9
26 : 0.91	30 : 0.9	21 : 1.19	14 : 1.22	32 : 0.91	5 : 0.89	29 : 1.22	24 : 0.9
32 : 0.84	14 : 0.83	29 : 1.14	20 : 1.21	14 : 0.9	22 : 0.83	32 : 1.07	31 : 0.84
14 : 0.8	33 : 0.79	22 : 1.02	16 : 1.2	33 : 0.89	29 : 0.79	22 : 0.96	6 : 0.8
16 : 0.78	32 : 0.76	30 : 0.83	30 : 0.84	30 : 0.8	30 : 0.72	30 : 0.88	30 : 0.77

IV. Model Training Information

Table IV.1. Parameter sweep possible and best parameters for Random Forest regression model

Model	Parameter sweep possibilities	Best parameters	Root mean squared error
Linear Regression	None	None	0.89
Random Forest Regressor fixed validation set	hyperparameters = 'n_estimators': [5, 11, 18, 25, 32, 38, 45, 52, 59, 66, 72, 79, 86, 93, 100], 'max_features': ['sqrt'], 'max_depth': [2,4,6,8,10,12], 'min_samples_split': [2,4,6], 'min_samples_leaf': [1,2,3], 'bootstrap': [True, False]	'n_estimators': 93, 'min_samples_split': 6, 'min_samples_leaf': 3, 'max_features': 'sqrt', 'max_depth': 8, 'bootstrap': True	0.8
Random Forest Regressor cross-validation	hyperparameters = 'n_estimators': [5, 11, 18, 25, 32, 38, 45, 52, 59, 66, 72, 79, 86, 93, 100], 'max_features': ['sqrt'], 'max_depth': [2,4,6,8,10,12], 'min_samples_split': [2,4,6], 'min_samples_leaf': [1,2,3], 'bootstrap': [True, False]	'n_estimators': 52, 'min_samples_split': 4, 'min_samples_leaf': 1, 'max_features': 'sqrt', 'max_depth': 8, 'bootstrap': True	0.8

The possibilities for hyperparameters were the same for all models:

- 'max_depth':[2,3,5,7,10,15],
- 'min_samples_leaf':[3,5,10,15,20],
- 'min_samples_split':[8,10,12,18,20,16],
- 'criterion':['gini','entropy']

Table IV.2. Parameter sweep possible and best parameters for Decision tree classifier models

Model	Best parameters	Accuracy
Decision tree classifier fixed validation set	'min_samples_split': 8, 'min_samples_leaf': 20, 'max_depth': 7, 'criterion': 'entropy'	0.8
Decision tree classifier with cross-validation no resampling	'min_samples_split': 12, 'min_samples_leaf': 3, 'max_depth': 3, 'criterion': 'entropy'	0.78
Decision tree classifier with cross-validation oversampled	'min_samples_split': 12, 'min_samples_leaf': 20, 'max_depth': 5, 'criterion': 'entropy'	0.78
Decision tree classifier with cross-validation undersampled	'min_samples_split': 16, 'min_samples_leaf': 15, 'max_depth': 5, 'criterion': 'entropy'	0.79
Decision tree classifier with cross-validation resampled using SMOTE-TOMEK	'min_samples_split': 12, 'min_samples_leaf': 20, 'max_depth': 5, 'criterion': 'entropy'	0.79

Table IV.3. How many times datapoints corresponding to counterfactuals were misclassified by the SMOTE-TOMEK model specified in table IV.2

Times misclassified	Question ID-s
6	7 ; 11 ; 12 ; 28
5	5 ; 6
4	10 ; 19 ; 20 ; 22 ; 23 ; 25 ; 29 ; 31
3	13 ; 17 ; 18 ; 26 ; 30 ; 34
2	8 ; 24
1	9 ; 15 ; 16 ; 21 ; 27
0	14 ; 32 ; 33

Table IV.4. How many times datapoints corresponding to individuals were misclassified by the SMOTE-TOMEK model specified in table IV.2.

Times misclassified	Person ID-s
11	113
10	91 ; 131 ; 181 ; 64
9	99
7	112
5	143
4	75 ; 1188
3	176 ; 108 ; 147
2	139 ; 79

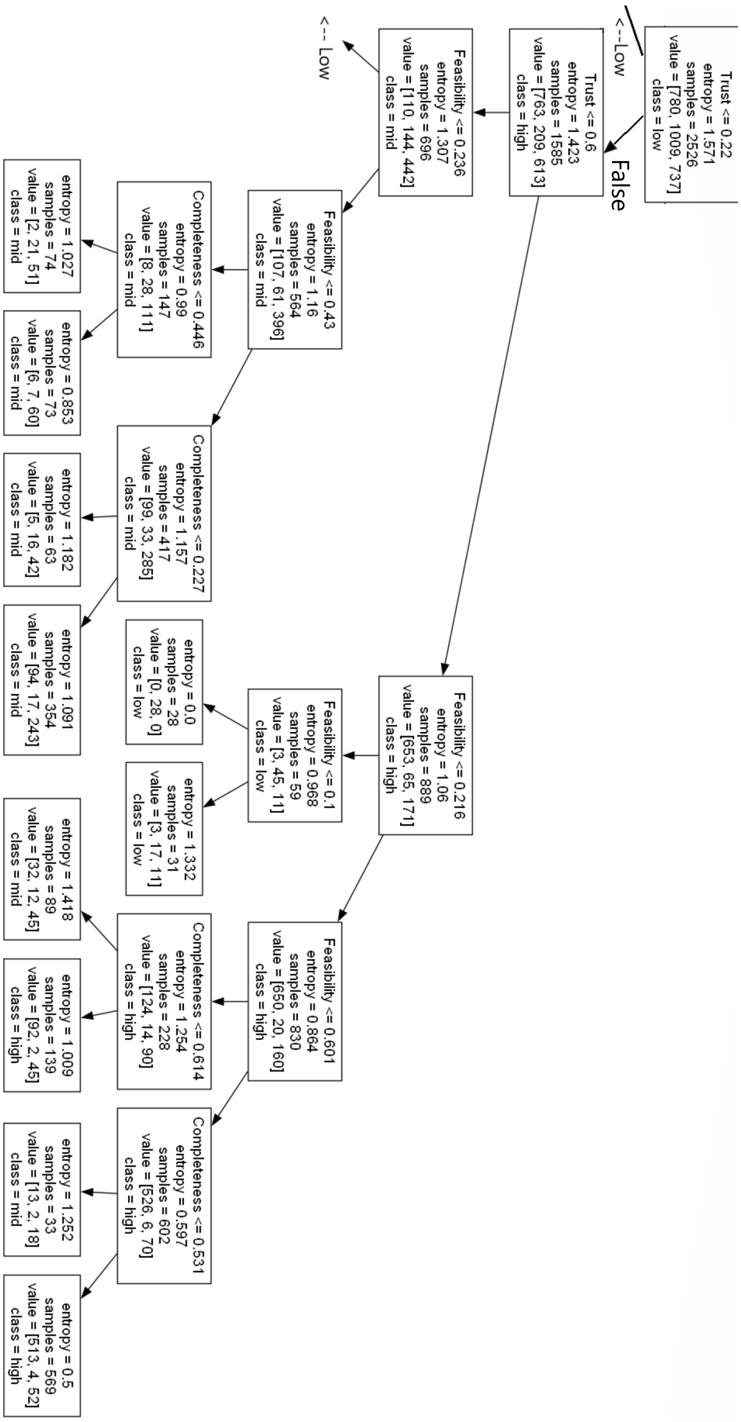


Figure IV.1. Decision tree visualisation example. Left half has not been added due to all the nodes leading to low values i.e. the first node at the top already sorts most of the low values with only Trust. The metric values used to split nodes have been normalised from a 6 point scale.

V. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Rasmus Moorits Veski**,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   Measuring Human Preferences in Counterfactual Explanations

   supervised by Marharyta Domnich, Kadi Tulver and Raul Vicente.

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 4.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Rasmus Moorits Veski
15.05.2024