Developing an Explainable AI Recommender System

Prabjot Kandola and Chris Baber

School of Computer Science, University of Birmingham

ABSTRACT

We used a theoretical framework of human-centred explainable artificial intelligence (XAI) as the basis for design of a recommender system. We evaluated the recommender through a user trial. Our primary measures were the degree to which users agreed with the recommendations and the degree to which user decisions changed following the interaction. We demonstrate that, interacting with the recommender system, resulted in users having a clearer understanding of the features that contribute to their decision (even if they did not always agree with the recommender system's decision or change the decision). We argue that the design illustrates the XAI framework and supports the proposal that explanation involves a two-stage dialogue.

KEYWORDS

Explainable AI, Recommender Systems, Travel planners

Introduction

Explainable AI (XAI) is a set of processes and methods to allow humans to comprehend the output of AI systems. Often these approaches emphasise the 'interpretability' of the model, i.e., how easily humans can understand the underlying model used by the AI system. An alternative approach is concerned with 'explainability' where the AI system is explaining its results, often in terms of the features which may have led to a particular output (Kaur et al., 2022; Erasmus et al., 2020). However, both approaches have a tendency to be AI-centric rather than human-centred, i.e., the approaches assume that the human needs to *understand* what the AI system has done and why it has done this. Such understanding need not be important to many forms of explanation (Mueller et al., 2019). Adadi and Berrada (2019) proposed four reasons as to why people need explanations from AI systems.

- Explain to justify: the AI system must justify why that explanation resulted;
- Explain to control: the AI system provides sufficient information for the user to identify and correct errors;
- Explain to improve: the user is able to correct the model that the AI system is using, so that the performance of the AI system can be improved;
- Explain to discover: the user is able discover the beliefs that the AI system is using, perhaps through testing with counter-factual examples.

When presented with XAI tools, there can be a tendency for users to over-trust such tools (Kaur et al., 2020) or the visualizations that are used (Hohman et al., 2020). Mueller et al. (2019) concluded that an explanation needs to focus on global rather than local explanations, on the performance of the user and encourage the user to reflect on their own interpretation of the output of the AI system.

In other words, the purpose of 'explanation' should not simply be to train the user to understand what the AI system is doing but to enable the user to better integrate the output of the AI system into their decision-making. This presents a departure from AI-centric approaches but faces two fundamental barriers:

(1) There are no universal criteria as to what defines an adequate explanation from an AI system. Therefore, AI system developers have no standard definition to follow when developing explanations;

(2) Even if there were universal criteria, these might not be applicable to users of the AI system for all contexts of use.

In previous work, we argued that an explanation ought not to be solely the concern of the direct user of the AI system but with anyone affected by the AI system, i.e., those who program the system as well as analysts who interpret its output and other stakeholders affected by the decisions based on the AI system's output. This is a tall order, but one that a human-centred approach could help address. To do this, we have proposed a framework that specifies the kind (s) of knowledge an AI system should provide so the 'Explanation' would be both 'interpretable' and 'explainable' to all stakeholders (either through their direct interaction with the AI system or through indirect actions, i.e., where the output of the AI system is communicated by an intermediary). More simply, XAI systems only focus on decision relevant features and the definition of 'relevance' that has been applied.

A model of Explanation

'Explanation' involves common ground in which two parties are able to align features to which they attend *and* the relevance that they apply to these features. We use this proposal as the basis for designing a recommender system. Baber et al. (2020, 2021) suggest that much of the prior work in XAI systems 'provides an output only at the level of features. From this the user has to infer Relevance by making assumptions as to the beliefs that could have led to that output. But, as the reasoning applied by the human is likely to differ from that of the AI system, such inference is not guaranteed to be an accurate reflection of how the AI system reached its decision.' That is, users are likely to have a different understanding (which may be due to demographics such as age, gender, education or salary) to the system in terms of 'why' a feature may have been chosen for a particular decision, or 'what' features could be used to reach a particular decision. Therefore, if the user disagrees with a decision, the user will have to infer what other features they would need to choose from to receive a recommendation which is more closely related to what they would want. In this case, the goal of explanation would be to align the features to which they attend and the type of relevance that they apply to these features. These assumptions are presented in figure 1. This framework suggests that an explanation involves an 'agreement on features (in data sets or a situation) to which the explainer and explainee attend and an agreement on why these features are relevant (this proposes three levels i.e., 'cluster' in which a group of features will typically occur together, a 'belief' which is a reason as to why these clusters occur, and policy which justifies the belief related to this action.

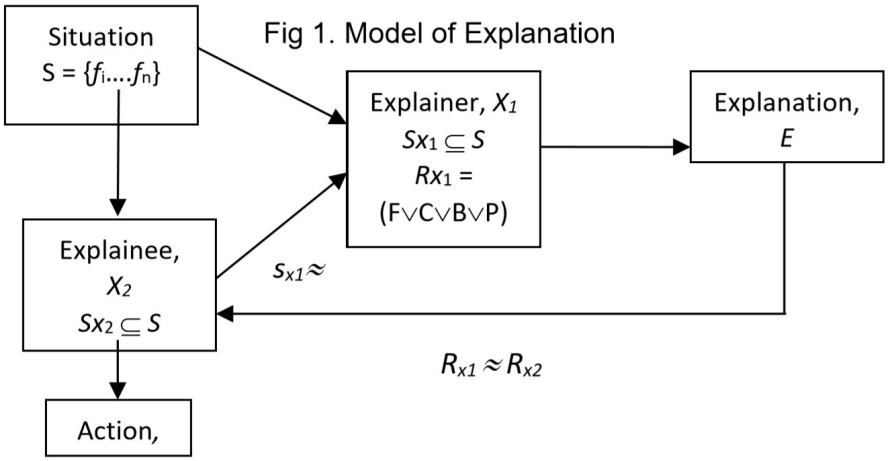

Figure 1: Framework for Human-centred XAI (Baber et al., 2021). A situation, S, has a set of features, {fi….fn}, which can be described symbolically, using words, numbers, pictures, etc. For example a situation might be the user choosing features (i.e. price, time) they believe are important when travelling from University of Birmingham to UoB. The Explainer is the XAI system's set of features which contribute to an Explanation. The Explainee is the users and includes the set of features to which they attend. Action is the action which could be taken by the user in light of the explanation.

An Explainable Journey Recommender System

In this paper, we develop a recommender system based on the model presented in figure 1There were two challenges in developing this recommendation system in order to meet the criteria suggested in the framework:

(1) The set of features which the Explainer X1 attends should overlap the set of features used by the Explainee X2

(2) Ensure the Explainer and Explainee can agree on what features define a situation i.e. define 'relevance'

The recommender system suggests routes for a user to take when travelling from University of Birmingham to the City Centre or vice versa. For the user to receive a recommendation they need to indicate features they believe are important when make a travel decision. Features could include 'price' or 'time', e.g., if the user chose 'time' and ranked this as '1', the recommendation would suggest taking an 'Uber' since it faster than Cycling or Walking.

The first challenge of the recommender system is for the Explainer (recommendation system) and Explainee (user) to attend to the same features. The second challenge of this system is for the Explainee and Explainer to have a similar concept of relevance which can achieved through dialogue between Explainer and Explainee (conducted, in this instance, using a chatbot).

Defining Features

The interaction commences with the user selecting a destination for the journey (figure 2). We use a user interface design familiar from ticket vending machines. The defines the scope of the Situation for the recommender system.

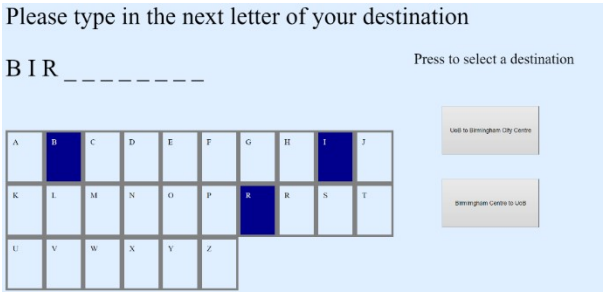

Figure 2: Defining a destination

As indicated in figure 1, a situation, S, has a set of features, {fi….fn}. In our model, we assume that the Situation also includes the constraints that define an acceptable decision. For this, we invite users to select 'features' that they believe to be relevant to their choice of journey type.

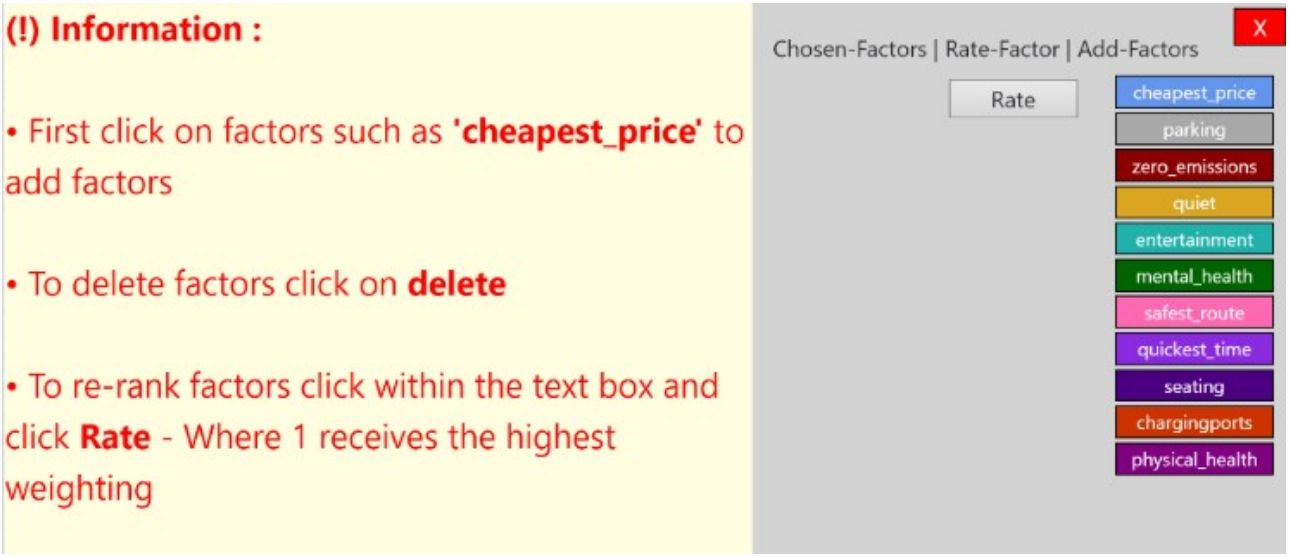

Figure 3: Defining Features

As figure 3 shows, users can select from a set of features (derived from an initial study with transport users). This initial set included {timing, price, emissions, congestion, capacity, number of changes, health, entertainment, charging ports, seating, safety, quiet, parking} can be expanded during user trials where participants offer additional features. The weighting of each selected feature is defined as a ratio of the number features selected such that the magnitude decreases, i.e., if the user selects 3 features then this produces weights of 0.5, 0.35, 0.15 (figure 4)

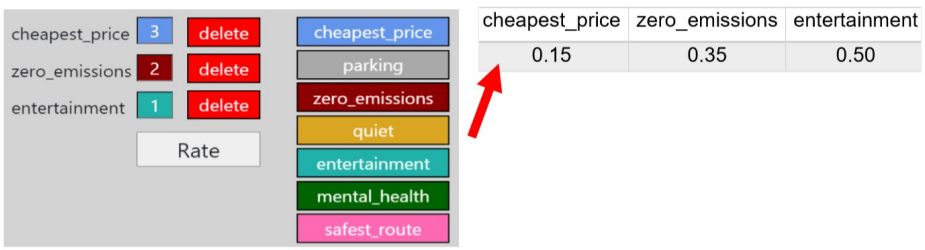

Figure 4: Calculating the weight of each feature

A pre-defined SQL database scores all features for each mode of transport {bus, taxi, car, train, walking, cycling). From this, the user weighting is combined with the mode of transport scoring. For example, the selected features map to the mode 'car' as shown in figure 5.

user_transport_weights1

cheapest_price	zero_emissions	entertainment
0.15	0.35	0.50

*

user_transport_weights1 (This is the car route)

cheapest_price	zero_emissions	entertainment
0.60	0.25	1.00

Figure 5: Ranking the features. In this case, the overall rating of mode: car is defined as: Price (0.15* 0.6) = 0.09 + Emissions (0.35*0.25) = 0.0875 + Entertainment (0.5*1) = 0.5 = 0.68.

Applying these features to the other modes of transport allows us to create a ranked list based on these ratings (figure 6).

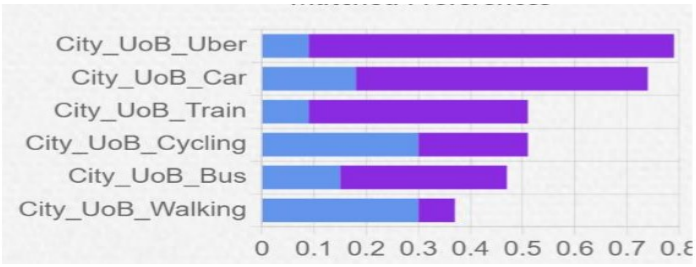

Figure 6: Rank ordering of all journey options

In addition to ranking transport relative to user-defined features, we also implement an algorithm which uses the features chosen by the user, as well as the features 'zero emissions' and 'physical health' to provide a recommendation. The 'zero emissions' and 'physical health' would always be given a higher ranking than the features selected by the user. While this process does *not* use Artificial Intelligence, we felt that it was sufficienty opaque for users to have difficulty in interpreting the recommendation and how it was derived. In other words, the purpose of this activity was not to simulate AI per se but to produce a recommendation that required explanation.

The recommender system implements a chatbot which makes the user aware of their features and how they relate to their routes as well as provides justifications into why specific decisions have been made to then nudge the user into changing their idea of what a good decision is based on the ideal recommendations (figure 7). This is where the system and the user work together to identify what features they believe are important and what a good decision is.

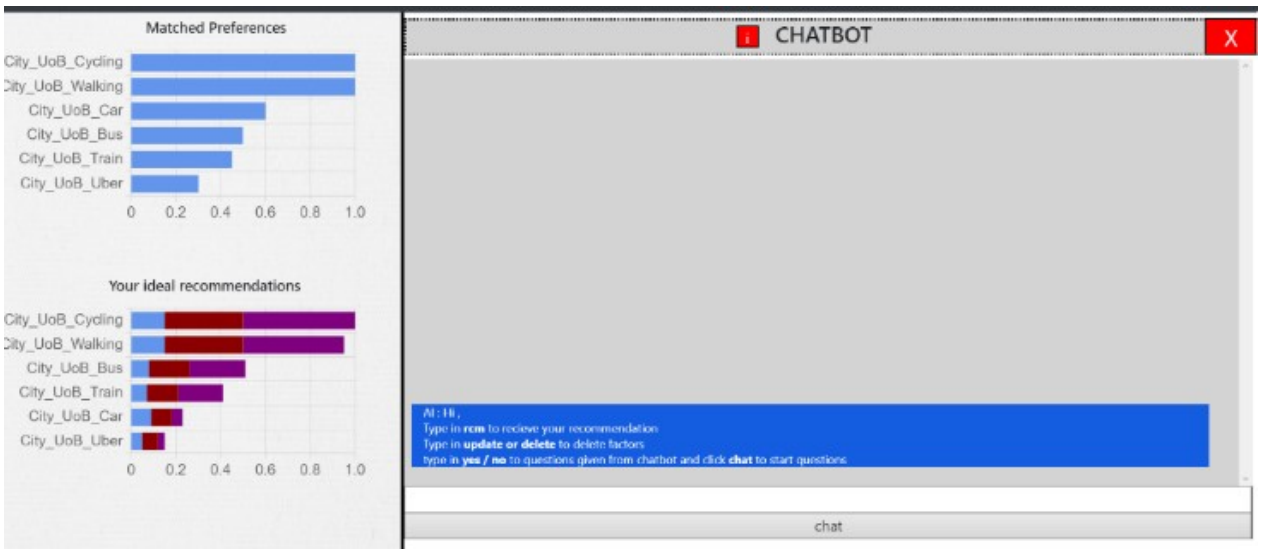

Figure 7: User interface showing preferences, 'ideal recommendations' and chatbot

Figure 7 shows the user interface with which the user can interact. There are three actions that the user can take: the user will agree with the explanation and choose a particular route; the user will disagree (not same) with the explanation, and 'Person 1' would be given new recommended features and then go to the action; the user can accept the recommendation and is shown a map with detailed instructions of the journey (figure 8).

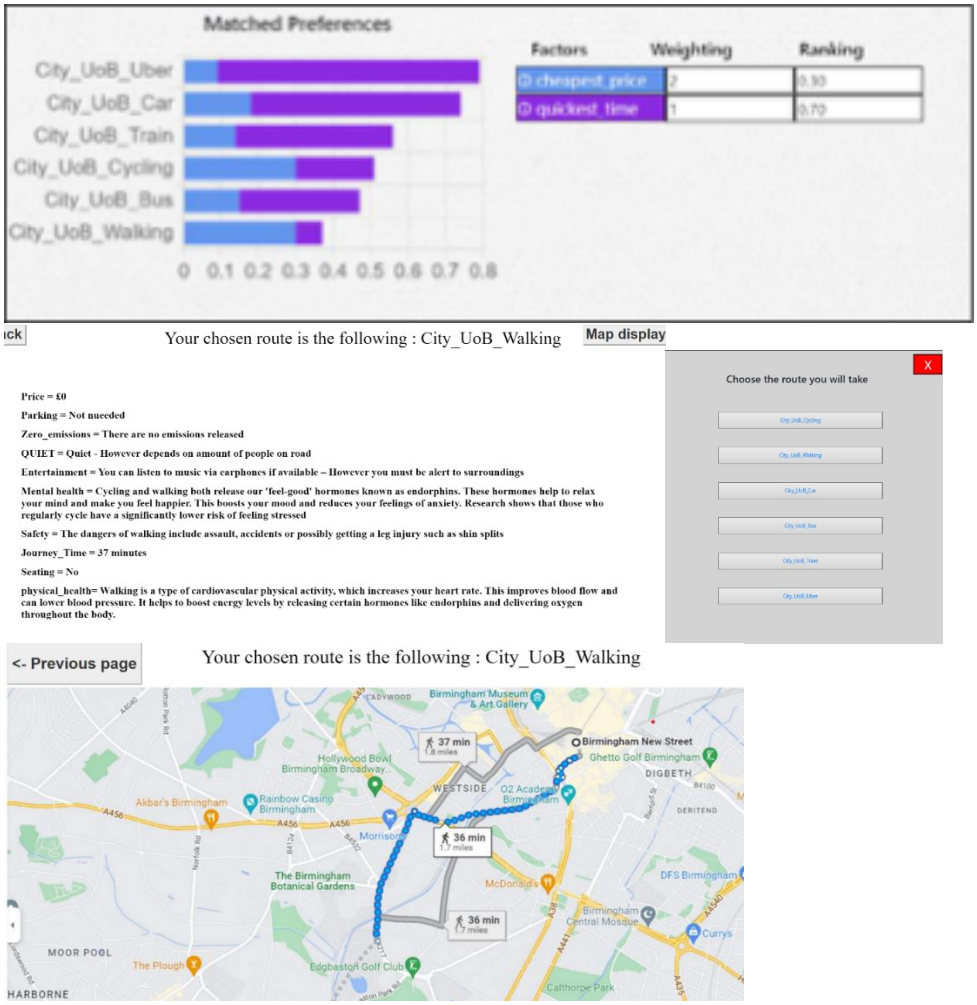

Figure 8: Journey plan output by the recommender system

Evaluation

20 participants were involved within this study, participants were either current or graduates from several universities. The age of participants ranged from 21 – 30. We accept that this produces a homogenous sample but propose that this makes it easier to aggregate the results. Future work could explore different user groups through more stratified sampling.

Participants were asked to interact with the recommender system in order to define a journey. They were asked, before the interaction began, what factors they normally consider when planning a journey. We used this to define the baseline against which we could compare the set of features that were considered following the interaction. As they interacted with the recommender system, we asked as a form of Cognitive Walkthrough them to articulate their impressions of the system's operation, whether they understood its recommendations, and whether the interaction had altered their choice of features or decision on journey type.

Prior to the interaction, the main features participants considered were 'price' and 'time'. This agrees with prior studies which states 'people typically only mention one or two features' [4]. For

participants, ‘time’ was associated with their experience of going into university such as arriving to a lecture on time or attending a meeting. Some users also noted ‘time’ as important since they don’t like to ‘waste time’ during the day, this could be because they have other activities such as the ‘gym’, ‘university work’ or wanting to go out with ‘friends and family’. Some users noted ‘weather’ as something they would consider; it was found that this would affect their travelling arrangements or the time they might leave their house. The type of transport they would take was also brought up, for example some users would often talk about ‘train’ or ‘uber.

Following the interaction, participants reported more features as relevant to their decision. Figure 9 shows the effect of interacting with the interactive chart or the chatbot on the number of features mentioned by participants.

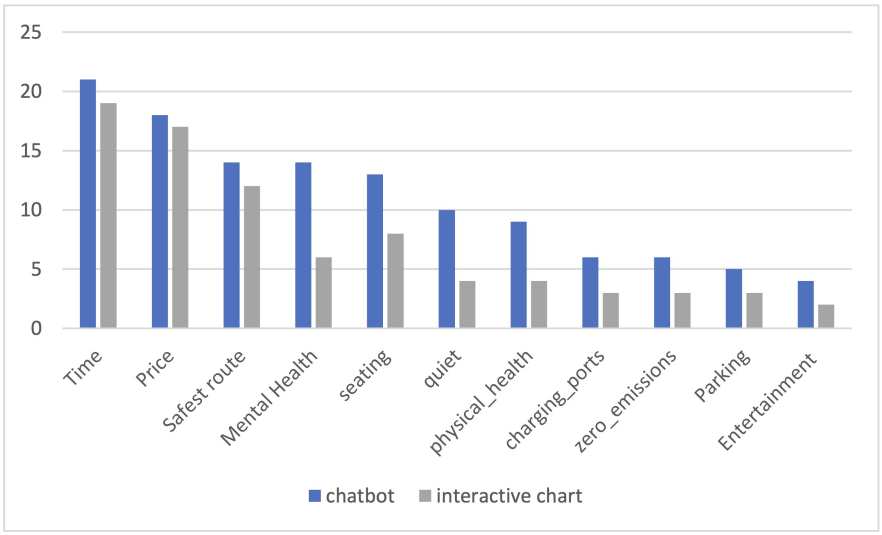

Figure 9: Count of features mentioned by participants

Combining participant response, we constructed a concept map(figure 10).

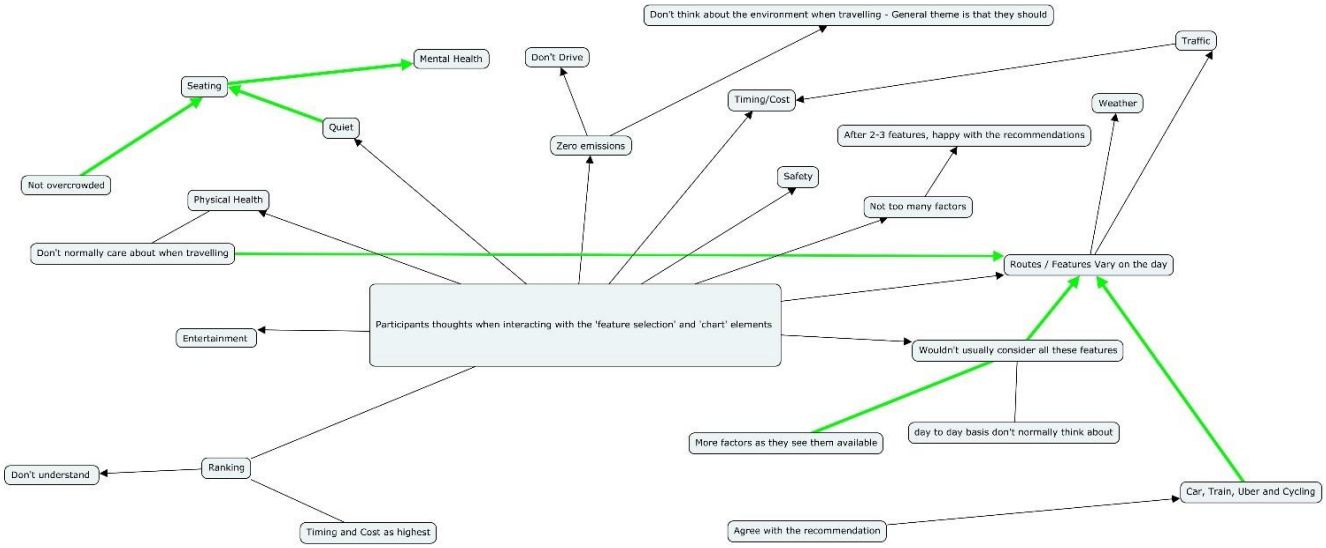

Figure 10: Concept map from participants following interaction with the recommender system

On whole, participants preferred their own ‘matched preferences’. However, the majority (17/20) of participants believed that the ‘ideal recommendations’ derived from physical health ‘made sense’ and these should be routes they should take. The reasons given for not including these features or taking the ideal routes was because they were ‘too lazy’, ‘cycling would take too long’ or ‘walking

would take too long'. Participants were less inclined to include zero emissions as a feature in their decisions (mainly as participants did not drive to university).

Furthermore, the chatbot helped participants understand 'why' features were chosen for their particular route, and once they understood the reasoning for this feature, they could then re-rank and alter this within their charts to receive a recommendation more closely related to what they would want, e.g., some participants did not understand how safety would lead to a higher rating for walking and thought a car to be safer. However, after understanding the XAI' s reasoning they agreed with the reasoning for this. When the chatbot gave its justifications for zero emissions it was found it was not enough to persuade the user into changing their minds, participants were hesitant to include zero emissions as they understood this would result in a change in the order of their recommendations where 'cycling' or 'walking' would be placed higher. This is preferable as participants now understood why and how features led to specific recommendations rather than having to 'inferred' this themselves.

To conclude, the recommendation system did not force users to change their minds or alter their choice. In this case, it was not particularly useful to 'nudge' their choices. However, this was not the primary intention of the project. Rather, we have demonstrated how a design for a recommender system can be developed from our XAI framework and that interacting with this recommender system helped users to elaborate on the features that inform their choice, and to understand how the recommender system has produced its recommendation – both of which we believe are integral to developing XAI.

References

Adadi, A. and Berrada, M., 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access, 6, 52138-52160

Baber, C., McCormick, E., & Apperley, I., 2021, A human-centered process model for explainable AI. In: Naturalistic decision making and resilience engineering symposium 2021, France.

Baber, C. McCormick, E. and Apperley. I., 2020, A framework for explainable AI, *Proceedings of Institute of Ergonomics and Human Factors*

Erasmus, A., Brunet, T.D. and Fisher, E., 2021. What is interpretability? *Philosophy & Technology*, *34*, 833-862.

Hohman, F., Wongsuphasawat, K., Kery, M. and Patel, K., 2020. Understanding and Visualizing Data Iteration in Machine Learning. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, New York: ACM, 1-13.

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H. and Wortman Vaughan, J., 2020, April. Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning, *Proceedings of the 2020 CHI conference on human factors in computing systems*, New York: ACM, 1-14.

Mueller, S.T., Veinott, E.S., Hoffman, R.R., Klein, G., Alam, L., Mamun, T. and Clancey, W.J., 2021. Principles of explanation in human-AI systems. *arXiv preprint arXiv:2102.04972*.

Mueller, S.T., Hoffman, R.R., Clancey, W., Emrey, A. and Klein, G., 2019. Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. *arXiv preprint arXiv:1902.01876*.