Human-centred explanations for artificial intelligence systems

Baber, C; Kandola, P; Apperly, I; McCormick, E

DOI:
10.1080/00140139.2024.2334427

License:
Creative Commons: Attribution (CC BY)

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Baber, C, Kandola, P, Apperly, I & McCormick, E 2024, 'Human-centred explanations for artificial intelligence systems', *Ergonomics*. https://doi.org/10.1080/00140139.2024.2334427

Taylor & Francis
Taylor & Francis Group

Ergonomics

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/terg20

Human-centred explanations for artificial intelligence systems

C. Baber, P. Kandola, I. Apperly & E. McCormick

ERGONOMICS
https://doi.org/10.1080/00140139.2024.2334427

Taylor & Francis
Taylor & Francis Group

ARTICLE

OPEN ACCESS

Human-centred explanations for artificial intelligence systems

C. Baber[a], P. Kandola[a], I. Apperly[b] and E. McCormick[b]

[a]School of Computer Science, University of Birmingham, Birmingham, UK; [b]School of Psychology, University of Birmingham, Birmingham, UK

ABSTRACT

As Artificial Intelligence (AI) systems increase in capability, so there are growing concerns over the ways in which the recommendations they provide can affect people's everyday life and decisions. The field of Explainable AI (XAI) aims to address such concerns but there is often a neglect of the human in this process. We present a formal definition of human-centred XAI and illustrate the application of this formalism to the design of a user interface. The user interface supports users in indicating their preferences relevant to a situation and to compare their preferences with those of a computer recommendation system. A user trial is conducted to evaluate the resulting user interface. From the user trial, we believe that users are able to appreciate how their preferences can influence computer recommendations, and how these might contrast with the preferences used by the computer. We provide guidelines of implementing human-centred XAI.

Practitioner summary: This paper presents a formal description of explanatory discourse for Human-Centred Explainable Artificial Intelligence (XAI) and demonstrate the use of this formalism as the basis for designing user interface for a recommender system. The recommender system is evaluated through a user trial. The paper concludes with guidelines for developing Human-Centred XAI.

ARTICLE HISTORY
Received 10 May 2023
Accepted 11 March 2024

KEYWORDS
Explainable artificial intelligence; user interface design; user trial

Introduction

Artificial Intelligence (AI) systems have demonstrated impressive performance, particularly in well-defined domains such as image processing or video-game playing. However, contemporary AI systems use techniques that can be opaque for the human user, which raises the challenge for AI systems to provide explanation (Neerincx et al. 2018; Rosenfeld and Richardson 2019) and there is growing requirement in Regulatory frameworks for AI to explainable, e.g. '*the development of intelligible AI systems is a fundamental necessity if AI is to become an integral and trusted tool in our society… in most cases we believe explainability will be a more useful approach for the citizen and the consumer.…*'[1]

Explainable AI (XAI) is a set of processes and methods intended to allow humans to comprehend the output of AI systems (Adadi and Berrada 2018). '*Given an audience, an explainable Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand*' (Barredo Arrieta et al. 2020, 6). As this quotation implies, the focus of XAI is on improving the ability of the AI system, rather than on ensuring that humans can fully understand the explanation as the basis for decision or action. There is growing recognition that such AI_centric approaches need to be challenged from a human-centred perspective (Hoffman, Klein, and Mueller 2018).

AI-centred (rather than human-centred) approaches assume that the purpose of explanation is to elucidate the mechanics through which a decision or recommendation was reached. However, this assumes that the human needs to *know* what the AI system has done and why it has done this (Springer 2019). Approaches which concentrate on explaining the algorithm might not be beneficial to users (Alufaisan et al. 2021; Bansal et al. 2021; Carton, Mei, and Resnick 2020; Edwards and Veale 2017; Ehrlich et al. 2011; Wang and Yin 2021). Understanding how a recommendation has been made need not be important to many forms of explanation (Mueller et al. 2019, 2021). When humans provide explanations to each other, they rarely engage

CONTACT C. Baber c.baber@bham.ac.uk School of Computer Science, University of Birmingham, Birmingham, UK

in detailed accounts of their reasoning processes (Klein et al. 2021). A more insidious problem (with the idea that explanation is about explicating the functioning of the AI system's decision) is that this implies persuasion, i.e. the AI system has produced an answer (whether or not the human agrees) and the role of explanation is to make the human accept this as the 'correct' answer. In human experience, an explanation is not necessarily intended to convince the recipient to change their mind; it is possible to appreciate the explainer's point-of-view and still disagrees (de Graaf and Malle 2017).

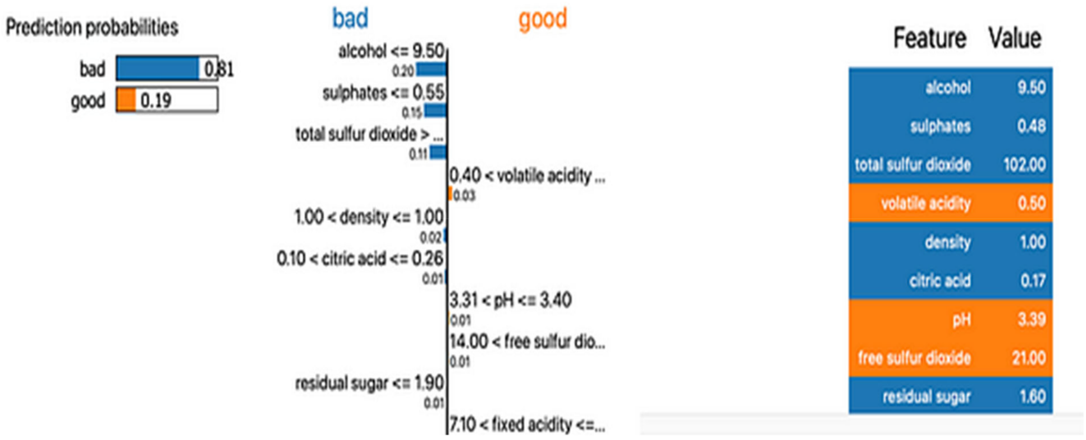
A common method in AI-centred approaches is the Local Interpretable Model-free Explanation (LIME) of Ribeiro, Singh, and Guestrin (2016). LIME presents the set of features that contribute to the algorithm's output. Figure 1 illustrates this with an example about features that affect wine quality and how these can be classified using a Random Forest Machine Learning classifier.

Figure 1 gives a score for the recommendation or prediction (on the left), an indication of which features contributed most strongly to the prediction (in the centre) and a table that reiterates the contributions (on the right). What LIME and similar approaches do not provide is an indication of *why* the algorithm has produced these scores, i.e. what assumptions, beliefs, expectations etc. have contributed to producing the score. In Figure 1, it would be necessary to not only appreciate the meaning of the output (which is to classify a particular bottle of wine as being 'good' or 'bad') but also an explanatory model that relates the features to this output. In the absence of the AI system providing an explanatory model, it is left to the human to infer or invent a model that they feel is plausible.

Mueller et al. (2019) concluded that an explanation needs to focus on global rather than local explanations, to consider the activity of the user, and to 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 in 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 *all* users of AI systems in *all* contexts of use.

Stuart Russell, in his 2021 Reith Lecture[2] on 'Living with AI', claimed that 'traditional AI' seeks to optimise a decision in terms of data and criteria but that 'future AI' should be designed to appreciate that humans might not know the exact criteria for a 'correct' decision or might not have clearly defined preferences. In this respect, 'future AI' should offer ways to help people ask better questions or better understand their own preferences (and the implications or trade-offs of combinations of preferences) or an appropriate explanatory model. The shift from finding patterns in data to



Feature	Value
alcohol	9.50
sulphates	0.48
total sulfur dioxide	102.00
volatile acidity	0.50
density	1.00
citric acid	0.17
pH	3.39
free sulfur dioxide	21.00
residual sugar	1.60

Figure 1. Example of LIME.[5] Figure 1 gives an example output of Local Interpretable Model-free Explanation (LIME) It uses an example about features that affect wine quality and how these can be classified using a Random Forest Machine Learning classifier. Gives a score for the recommendation or prediction (on the left), an indication of which features contributed most strongly to the prediction (in the centre) and a table that reiterates the contributions (on the right).

finding questions to ask, requires AI systems to reason about their own reasoning and decision-making (as well as being able to consider how the users of the AI system will reason and make decisions). Rather than blandly presenting an 'answer' or the features they use, AI systems ought to be able to discuss options available to their human users (with the AI system predicting the likely consequences of different options). In this way, explanation is not the account of how the answer was produced, but a conversation about how different answers reflect different preferences, different outcomes, and different explanatory models. We term this an explanatory discourse and, in the next sections, formalise this.

Explanatory discourses

In an early attempt at a formal definition of explanation, Hempel (1924) proposed a 'Covering Law Model' of History. A core question for historians is *why* a given Event occurred. Hempel suggested that a set of prior events could be regarded as antecedent Causes, combined according to some 'Law'. From this, an argument could be presented (either deductively or inductively) that the occurrence of antecedents increases the probability of the Event occurring. This suggests that the explanatory discourse (between two historians) would involve the statement of Causes to explain an Event. But the approach collapses under counter-examples (Salmon 1998) and is seldom espoused or defended nowadays. Hempel's argument relies on a 'common-sense psychology' that an explanation involves advancing an hypothesis through which events can be explained by contributory factors. This is similar to the way in which LIME (Figure 1) displays the features that contribute to an outcome of the algorithm. While contributory factors might superficially capture how we reason about events, they do not tell us whether the hypothesis is correct (or even testable), or whether the selection of contributory factors is complete or relevant. A second problem with Hempel's approach is that it defines an explanation as a casual model of an event that would be correct if the set of contributing factors *were* to occur, rather than showing *how* the contributing factors relate to each other.

While Hempel's approach has problems, the idea of creating a formal description of explanation has been attractive for XAI. Rosenfeld and Richardson (2019) defined explanation in terms of the interpretability (by a human) of the relationship between a target output, *T*, of an algorithm, $\mathcal{L}$, and the specific features, *F*, in a record of data, *R*, as: $Explanation = \mathrm{I}(\mathcal{L}(R\,x\,F, T))$

In a similar vein, Holzinger, Carrington, and Müller (2020) propose the System Causability Scale which suggests that aspects or features of a situation are combined into the explanation. In this the human or machine produces a statement, *s*, which is a function, *f*, of contributing factors such that $s = f(r, k, c)$, where *r* is the representation of unknown fact relating to an entity; *k* is pre-existing knowledge; c is the context in which an explanation is presented. Holzinger, Carrington, and Müller (2020) assume that human and AI system have equal access to a 'ground truth'. From this, explainability '*…highlights decision relevant parts of machine representations…, i.e. parts which contributed to model accuracy in training or to a specific prediction.* [Holzinger, Carrington, and Müller, 2020 195]'. This feels similar to Hempel's Covering Law Model, and implies 'ground truth' (i.e. the relationship between features and situation) can be fully defined. But, simply stating the features without an indication of why these (rather than other features) were selected might not lead to a useful or usable explanation. Implicit in this approach is the further assumption that the AI system's reasoning can be 'surfaced' (i.e. brought to awareness and expressed in words). AI systems might be unable to introspect on their own processes. But surfacing presents problems for humans as well because it requires us to introspect on our cognitive processes but also to put the tacit knowledge that this implies into words.

Langley (2019) defines an agent capable of producing an explanation as acting as follows:

> Given: Knowledge defining a space of possible solutions;
>
> Given: Criteria for evaluating candidate solutions;
>
> Given: An annotated search tree that includes solutions for some reasoning task…;
>
> Given: A query about why a solution ranks above others;
>
> Produce: An explanation why the solution is preferable.

Langley (2019) seems to assume that 'explanation' means acceptance by the user. So, from this definition, explanation cannot be challenged. One might argue that each of the 'givens' in the above definition could be individually challenged, but there is no obvious process inherent in this definition that modifies the explanation that is produced. In other words, this definition rests on the assumption of transmission of the explanation to the explainee. Further, while the definition of situation that humans create might be causal (e.g.

based on plausible 'causes' of a given event or feature), it is more likely that the definitions machines create are relational (e.g. based on correlation, regression, distance, similarity, etc.). This leads to the problem of mistaking correlation for causation, i.e. the human could misinterpret correlations, on which the AI systems depend, for either causal (i.e. generalisable) relations or predictive beliefs. But neither of these (causal relations or predictive beliefs) are integral to the AI system.

As Miller (2019) notes, a problem with an explanation that presents outcome plus features, is that ultimately these are based on the algorithm designer's intuition of what makes a 'good' explanation rather than on a sound understanding of how humans respond to, and make use of, explanation. This does not indicate why *some* features were selected or why the recommendation is appropriate to the user's concerns. Hoffman, Klein, and Mueller (2018) provide a comprehensive review of literature relating to explanation and make a convincing argument that explanation involves sensemaking by the human (to contextualise the output of the AI system). Sensemaking relies on the recognition that the process (of providing and receiving an explanation) must be reciprocal, iterative, and negotiated. In other words, rather than the human merely as the passive recipient of the AI system's explanation, there is a need for this to be a process through which an explanation is constructed through explanatory discourse. Explanations between humans recognise this problem and explanatory discourse but these techniques have not been commonly applied to XAI (Miller 2019).

HXAI: human-centred framework for XAI

Our aim is to produce a formal description that can reflect different types of explanatory discourse, that is applicable to human-human conversation and human-AI interaction, and that allows us to ask *how* explanations are produced. Initial versions of this framework have introduced a formal description and provided examples (Baber, McCormick, and Apperley,2020 2021). Maathuis (2023) comments positively on our approach which she describes as a 'formalism containing a situation explained by an explainee through an explainer by producing a corresponding situation based on a relevant action' but notes that our process model is, to date, qualitative. In this paper, we use a simple design and evaluation exercise to further illustrate our approach.

Figure 2 indicates that an explanation, *E*, occurs in, and relates to, a situation, *S*, which has a set of features, $\{f_i \ldots f_n\}$, that can be described symbolically, using words, numbers, pictures, etc. In this respect, a set of features could be analogous to the data which contribute to a frame in Klein et al. (2007) Data Frame Model. However, both 'data' and 'frame' have privileged meanings in the AI literature, so to avoid confusion we use the term 'situation'. This has the advantage, for a Human Factors audience, of calling to mind Situation Awareness, particularly when this is Distributed (Stanton et al. 2006) between agents. A 'feature' is some aspect of the situation to which people can attend and individuals in a situation ground their Situation Awareness, s_i, by attending to a subset of all features in *S*, i.e. $s_i \subseteq S_i$. For Distributed Situation Awareness, an important step is to establish common ground (Clark 1991; Clark and Brennan 1991) through which the situation can be agreed because features are external to individuals, in that anyone in *S* ought to be able to attend to the same features. In LIME, the features that the computer uses are presented to the user, but there is no scope for the user to offer the features they prefer or to challenge the ones offered by the computer.

A first challenge in producing an explanation is to ensure that the features to which the Explainer, X_1,

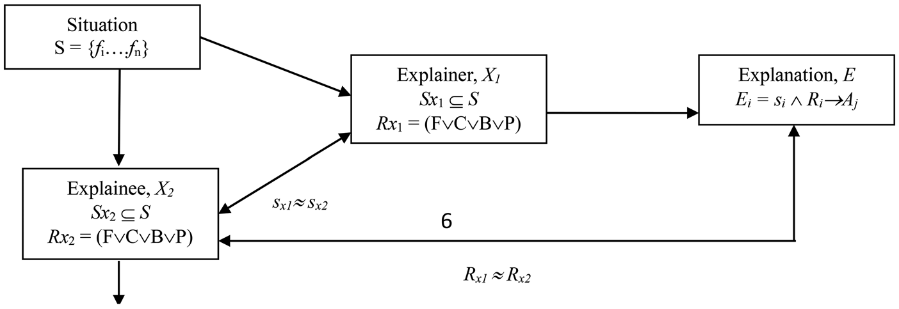


Figure 2. HAXI framework. This is a figure with text-boxes connected by arrows. In the top left of the figure there is a box labelled 'Situation'. This is connected to a box labelled 'Explainee' (directly below) and to a box labelled 'Explainer' (diagonally to the left). There is also a double-headed arrow connecting the 'Explainee' and 'Explainer' boxes. To the right of the 'Explainer' box is a box labelled 'Explanation' and there is an arrow from the 'Explainer' box to the 'Explanation' box. There is an arrow from the 'Explanation' to the 'Explainee' box. Finally, there is an arrow from the 'Explainee' box to a box labelled 'Action', directly below it. Each box contains annotations that are provided in the main body text.

attends will overlap with the set of features used by the Explainee, X_2. The definition of features for a situation will reflect the familiarity of the explainer and explainee with the situation, and their knowledge, expertise, judgement, and ability.

In human interactions with other people, we tend to offer one or two features as first-pass explanation (McClure et al. 2001; Leddo, Abelson, and Gross 1984; Tversky and Kahneman 1983). These features imply (a) a string of causal reasoning that the other people are assumed to be able to perform (i.e. we assume that, in terms of prior knowledge, $X_1 \approx X_2$), and (b) to be sufficient to explain the situation. We assume 'honest signalling' (Maynard Smith and Harper 2003) in that the feature is relevant to the situation. If there is a mismatch between selected features, e.g. as indicated by the explainee appearing puzzled or asking questions, then a further step will be required to align the selected features. This raises the next challenge which is to agree *why* specific features relate to the Situation, i.e. to define 'relevance'. Relevance, R, can be defined in terms of:

- *Features, F:* features in the situation to which both parties can attend (as indicated by $s_{x1} \approx s_{x2}$ and the common ground this implies);
- *Clusters*, *C*: features which typically co-occur in similar situations (and which can be used to predict likely outcomes in familiar situations);
- *Beliefs*, *B*: the reason why clusters co-occur, and which can predict consequences if specific features alter, and which allows inferences about causality to be made;
- *Policies*, *P*: rules which allow actions to be linked to clusters or features.

From this, an Explanation, *E,* involves the set of Feature, {f_i….f_n}, to which a person attends in a situation, S, in terms of the relevance, *R*, and a (potential) aim of influencing Action, *A*:

$$E_i = s_i \wedge R_i \rightarrow A_j, \quad \text{where } \mathrm{R} = (\mathrm{F} \vee \mathrm{C} \vee \mathrm{B} \vee \mathrm{P}) \qquad [1]$$

Figure 2 suggests that explanation involves checking the features attended by x_1 and x_2. If these differ, then the first-pass Explanation might involve highlighting specific features, so that $s_{x1} \approx s_{x2}$. Where there continues to be uncertainty or disagreement in the conversation between explainer and explainee, then further action might be required to produce agreement across one or more type of Relevance. Misalignment of Belief could involve challenging the selection of features; misalignment of Cluster could involve analysis using a different set of features; misalignment of Policy could involve proposing a different action. Of central importance to this process of explanation is the ongoing dialogue between explainer and explainee. Relating Figure 2 to the previous discussion on explanatory discourse, we assume four types of explanatory discourse in which the definition of the situation or Relevance are Aligned or Challenged (Table 1).

Designing a recommender system using the HXAI framework

In this section, we present a recommender system based on the model presented in Figure 2.

A prototype recommender system is developed to suggest 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 factors that influence their preference for a travel decision. Figure 3 shows the sequence of screens with which the user will interact when using the prototype recommender system. A justification for each screen is presented, in terms of the HXAI framework (Figure 2), in the following discussion.

Defining a situation (S1 ≈ S2)

The explainer (Recommender System) and explainee (user) will define the situation. As indicated in Figure 2, a situation, *S*, has a set of features, *{fi….fn}*. In this example, the features of the situation are:

i. The destination to which the user intends to travel,
ii. The factors that influence the user's preference for a mode of transport.

The interaction commences with the user selecting a destination for the journey (Figure 3, step 1). Here, the user interface is familiar from ticket vending machines.

Table 1. Types of explanatory discourse.

	Align	Challenge
Definition of Situation	Explainer draws attention to specific features in the situation.	Explainee disputes the indicated features and requires clarification of the theory or model being applied.
Definition of Relevance	Explainer presents underlying rationale for the theory or model.	Explainee offers alternative definitions of relevance or appeals to 'counter-factual' (or 'what if') examples, e.g. what if a given feature was present or absent.

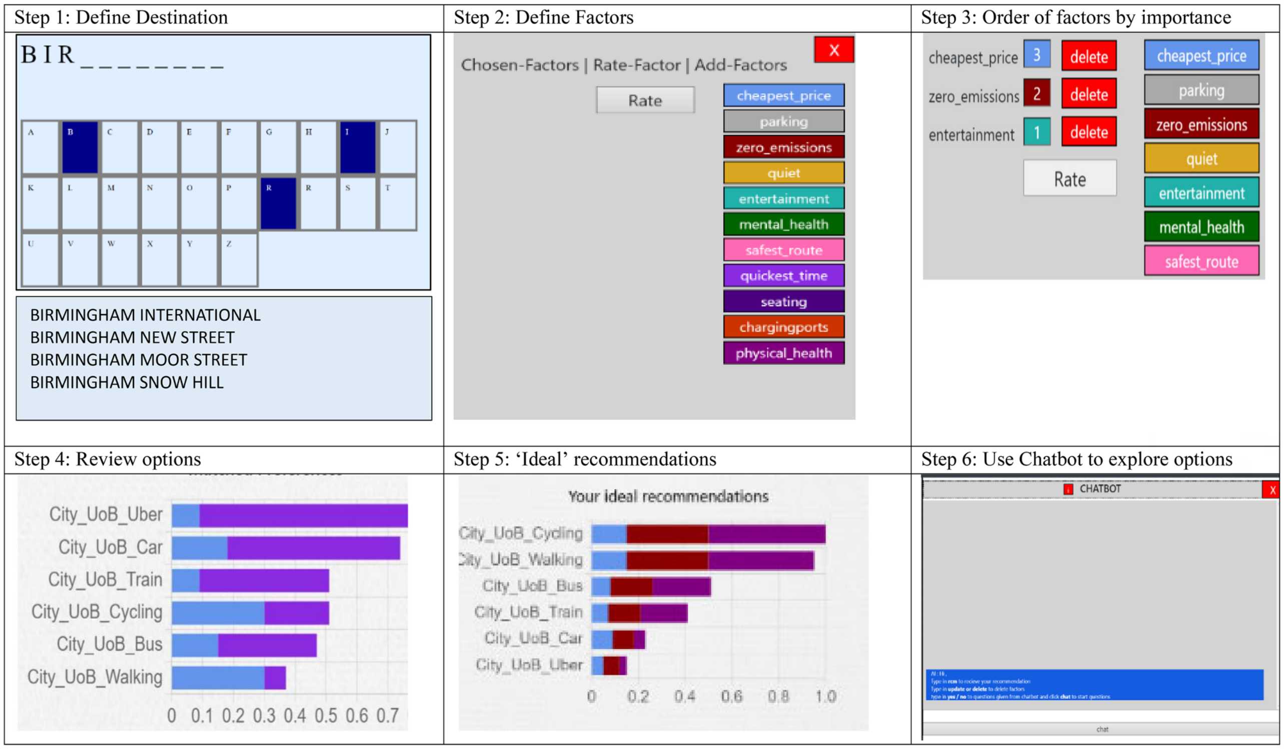


Figure 3. Screens from Recommender System. This shows six images, arranged in 2 rows of 3, from the user interface of the Recommender System. These are arranged in steps from 1 to 6. Step 1: Define destination, first image on top row, shows a user interface of a ticket vending machine with a keyboard laid out in alphabetical order, and the letters 'B, I, R' in a selection panel. Step 2: Define factors, middle image on top row, has a list of factors that could affect a person's choice of journey. Each factor is a label for a button and the buttons are arranged in a list. The factors are: cheapest price, parking, zero_emissions, quiet, entertainment, mental_health, safest_route, quickest_time, seating, charging ports, physical health. When a user clicks one of these buttons, the factor is added to the 'chosen factors' list (on the left of this screen). In the centre of the screen is a button labelled 'Rate' (which the user presses to arrange the chosen factors in order of preference). Step 3: Order of factors by importance, right of top row. This shows the chosen factors in order and their importance. In this case, the chosen factors are cheapest_price, zero_emissions, entertainment. Step 4: Review options, left on bottom row, shows a stacked bar chart with the possible routes order by score from the Recommender System. Each bar has two colours (indicating contribution of two factors to the recommendation). City_UoB_Uber has a score of 0.7; City_UoB_car has a score of 0.68, and other options have scores of 0.5 or less. Step 5: 'Ideal recommendations, middle of bottom row, shows the factors calculated by the recommender system as an 'ideal' solution in a stacked bar chart. This has 'City_UoB_Cycling 1.0 and City_UoB_Walking 0.9 as the top two recommendations. Step 6: Use chatbot to explore options, bottom right. This shows the first screen for the chatbot, inviting the user to ask a question.

This provides an initial feature to define the Situation. In addition to destination, we assume that the Situation also includes user's preferences for mode of transport. For this, we invite users to select factors that they believe to be relevant to their choice. As Figure 3, step 2 shows, users can select from a set of factors. This set was defined using a focus group of five people who regularly commuted between the University of Birmingham and the City Centre, and consists of {timing, price, emissions, congestion, capacity, number of changes, health, entertainment, charging ports, seating, safety, quiet, parking}. We propose that asking users to select the factors will support the explanatory dialogue around aligning the definition of a situation in Table 1 through encouraging users to define their preferences.

Once a factor has been selected, the user is invited to rank this relative to other factors that they have selected. Each factor is weighted by this ranking such that there is as equal an interval as possible between items in a set so that the first item in the ranking will have a higher weight than the second etc., e.g. if the user selects three factors then this produces weights of 0.5, 0.35, 0.15, i.e. assigning 0.5 to the first item, will leave 0.5 to be shared between the next two items and, keeping the interval between these as similar as possible, we have 0.35 and then 0.15. Weighting the factors will support the explanatory dialogue around aligning the definition of relevance in Table 1. While we invite users to indicate their preferences, this is not to assume that users will always have a clear idea of what their preferences are or be correct in their selection of preferences (Krishna et al. 2022). However, we believe that asking users to select and weight preferences in this manner helps to 'surface' these (as discussed earlier) and comparing these with preferences presented by a computer could help people reflect on their choices and the implications of these choices.

To simulate a computer generated recommendation, an SQL database of scores for all factors for each mode of transport {bus, taxi, car, train, walking, cycling) as they relate to the journey was created. From this, the user weighting is combined with the mode of transport scoring. This is used to vary the scores for the factors. For example, 'parking' has a different score for 'city centre' (where there are several car-parks with different prices and a congestion charge for some vehicles) than 'University' which has fewer car-parks (which are often full) (Table 2).

The selected features from Figure 3, step 3 map to the mode 'car' as shown in Table 5. Applying these features to the other modes of transport allows us to create a ranked list based on these ratings. Figure 3, step 4 presents the weighted features in a stacked bar graph. This is intended to provide the features that have contributed to the recommendation in much the same way that LIME, Figure 1, presents features to users.

In addition to ranking mode of transport relative to user-selected factors, we add the factors 'zero emissions' and 'physical health'. The computer would always rank 'zero emissions' and 'physical health' higher than the factors selected by the user. The rationale for this was the Recommender System might be seeking to 'nudge' the user into changing their preference for a mode of transport. This is shown, in Figure 3, step 5, as 'Your ideal recommendations'. The intention behind these factors is to generate recommendations that might challenge the user, or which are not immediately obvious to the user. For this study we present users with a recommendation that they would not be able to calculate without knowledge of the algorithm that generated it. While this process does *not* use Artificial Intelligence or Machine Learning, we felt that it was sufficienty opaque for users to have difficulty in interpreting the recommendation. In other words, the purpose of this activity was not to simulate AI systems *per se* but to produce a recommendation that required explanation. There are two actions that the user can take: the user can accept the recommendation and is shown a map with detailed instructions of the journey (Figure 4), or the user can disagree with the recommendation and seek further explanation by defining relevance through discussion with a chatbot.

Defining relevance (R1 ≈ R2)

The second challenge of this system is for the explainee and explainer to define relevance. The recommender system uses a chatbot to indicate how user-selected factors relate to their mode of transport (Figure 3, step 6).

Allowing users to ask 'why' when interacting with a movie recommender chatbot (Wilkinson et al. 2021) was shown to have positive benefits on user experience. However, this did not allow a dialogue between user and chatbot to refine the criteria for the recommendation.

In this prototype, the chatbot is implemented using JavaScript. This operates in an HTML page that has text fields for input (from the user) and output (from the chatbot). A set of arrays were predefined that related to specific types of user input. The types of utterance were constrained to only include comments or questions relating to the factors that were defined in Figure 3. These constitute the triggers to which the chatbot responds. For example, the user could ask 'why do you recommend<option>?' or 'why is<factor$_i$> rated higher than<factorj>?' While this produces a restricted dialogue, we felt that it was sufficient to simulate explanatory dialogue and to encourage user interaction. If users do not accept the chatbot's explanation, they can ask more questions to challenge the chatbot or get further information. Alternatively, the user can reconsider the features and their weighting (Figure 3, step 3) to revise the importance that they give to specific features. This can result in a change in the options (Figure 3, step 4) or the definition of 'ideal' recommendations (Figure 3, step 5). This is intended to support the explanatory dialogue relating to challenging the definition of relevance (Table 1).

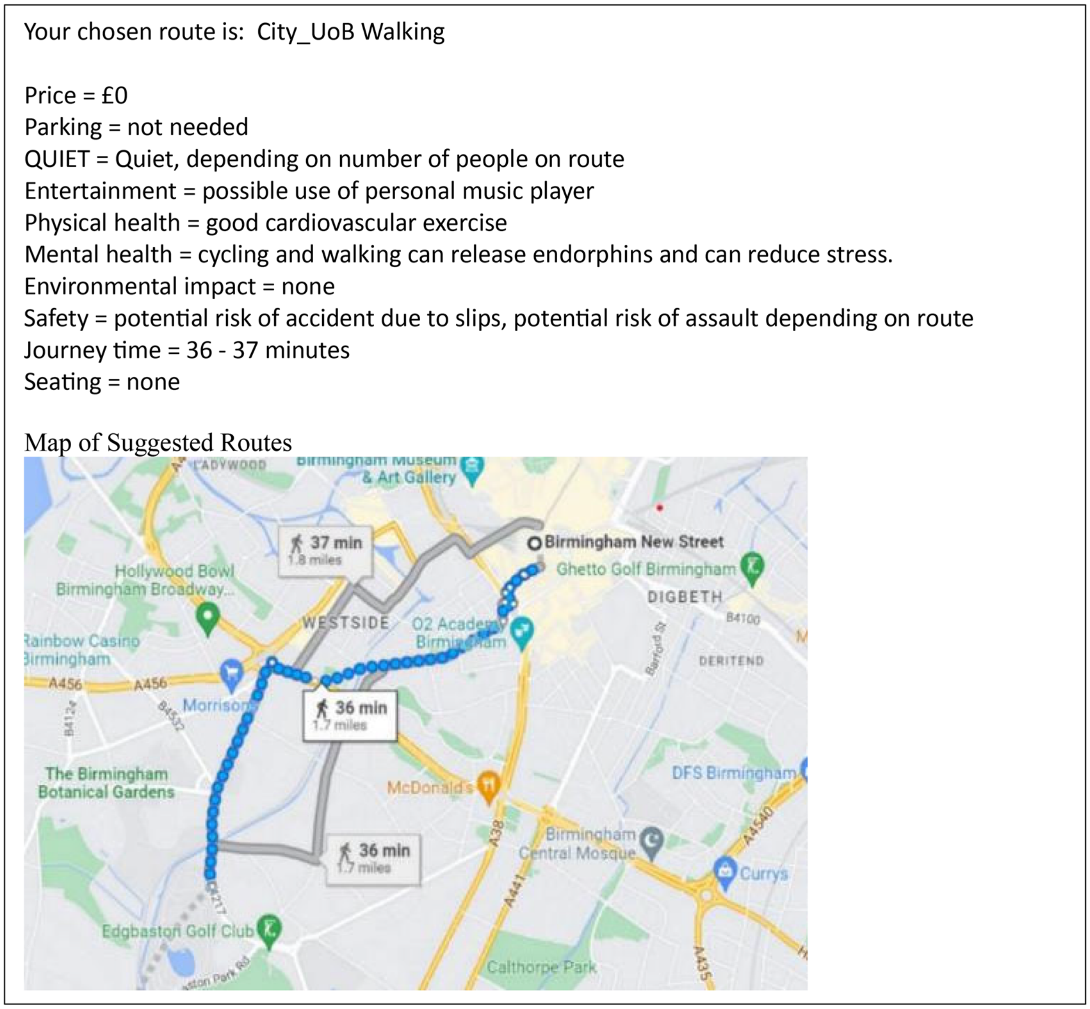
Producing a recommendation

If the user accepts the recommendation, this can be displayed in detail with a map showing the route (Figure 4). Relating the Recommender System design to Figure 2, we have made several design decisions which are summarised in Table 3. The first four rows in Table 3 highlight points already discussed. The final

Table 2. Ranking the factors.

Weights from user ranking				Scores for 'Car', journey to 'city centre'		
Cheapest_price	Zero_emissions	Entertainment	*	Cheapest_price	Zero_emissions	Entertainment
0.15	0.35	0.5		0.7	0.4	1.0

In this case, the overall rating of mode: car is defined as: Price (0.15* 0.7) = 0.105 + Emissions (0.35*0.4) = 0.14 + Entertainment (0.5*1) = 0.5 = 0.745.



Your chosen route is: City_UoB Walking

Price = £0
Parking = not needed
QUIET = Quiet, depending on number of people on route
Entertainment = possible use of personal music player
Physical health = good cardiovascular exercise
Mental health = cycling and walking can release endorphins and can reduce stress.
Environmental impact = none
Safety = potential risk of accident due to slips, potential risk of assault depending on route
Journey time = 36 - 37 minutes
Seating = none

Map of Suggested Routes

Figure 4. Journey plan output by the recommender system. This shows three panels. Top left is a panel that summarises the mode of transport that corresponds to the user's selected factors. To the right of this is a column of buttons that allow the user to select a different mode of transport. Below these is a map showing the route and journey time (taken from GoogleMaps).

Table 3. Summarising the design concept using the H-XAI framework.

Definition of situation	Definition of relevance	Explanatory discourse	Explainee's expected outcome	Example
Similar $S_{x1} \approx S_{x2]}$	Similar $R_{x1} \approx R_{x2}$	Align	Agreed response to situation	The recommended route is the one the user has chosen.
Different $S_{x1} \neq S_{x2}$	Different $R_{x1} \neq R_{x2}$	Challenge	Agreed definition of situation	The recommender defines an option using criteria the user had not previously considered.
Similar $S_{x1} \approx S_{x2}$	Different $R_{x1} \neq R_{x2}$	Challenge	Agreed definition of relevance $R_{x1} \approx R_{x2}$	The user does not agree with recommender's criteria for defining options. In this case, the user can change the weighting of options (Figure 4, step 3) or can use the chatbot (Figure 4, step 6) to accept or reject the criteria proposed by the recommender.
Similar $S_{x1} \approx S_{x2}$	Different $R_{x1} \neq R_{x2}$	Align	X_2's definition of relevance matches a subset of X_1's $\Delta R_{x2} \approx r_{x1} \subseteq R_{x1}$	The recommender introduces additional information to the user that can change their definition of relevance.
Different $S_{x1} \neq S_{x2}$	Different $R_{x1} \neq R_{x2}$	Align	$\Delta R_2 \approx r_1 \subseteq R_1$	The recommender could seek to 'nudge' (Caraban et al. 2019) through 'choice architectures' that present alternative actions in ways that are intended to support positive changes in behaviour. These technologies encourage or discourage behaviours**, i.e.** $A_2 = \Delta s_{2.}$There is no implication that the human needs to understand why this action has been proposed.
Different $S_{x1} \neq S_{x2}$	Similar $R_{x1} \approx R_{x2}$	Align	Align	The chatbot (Figure 4, step 5) has a model of reasoning towards conclusions (arguments). Through argumentation, parties identify points of similarity and difference, e.g. features to emphasise or notion of relevance. The user could then *explore* the effect of adding or removing features or changing relations, which could be particularly useful for counter-factual reasoning (Guidotti et al. 2019).

two rows in Table 3 indicate features that one might expect of a recommender system, i.e. nudging or argumentation. Where there is an option for users to change or retain their definition of situation or relevance (Table 3, rows 2 and 4), we might assume that this would be learned by the AI system, which could update its model of the user's preferences (although this is not implemented). We are interested in how interacting with the computer could encourage participants to reflect on their preferences for factors and to understand why the computer was making its recommendations.

Evaluating the recommender system

Producing a User Interface that supports explanation begins with the appropriate model of explanatory discourse that you are seeking to support. Additionally, the design should be reviewed and critiqued in terms of its potential to support explanation. There are several approaches to defining the quality of explanations (Schwalbe and Finzel 2023). Our preference is for Hoffman et al.'s. (2018) 'Explanation Goodness Checklist' and we have adopted this in our design process as an initial 'sanity check' of the design concept.

The terms used in the checklist are listed below, but we advise the reader to consult the original source for the checklist and its derivation (Hoffman et al. 2018):

- The explanation helps me understand how the [software, algorithm, tool] works.
- The explanation of how the [software, algorithm, tool] works is satisfying.
- The explanation of the [software, algorithm, tool] sufficiently detailed.
- The explanation of how the [software, algorithm, tool] works is sufficiently complete.
- The explanation is actionable, that is, it helps me know how to use the [software, algorithm, tool]
- The explanation lets me know how accurate or reliable the [software, algorithm] is.

The explanation lets me know how trustworthy the [software, algorithm, tool] is.

Table 4. Applying the explainability checklist to the recommendation system.

Explainability checklist	Design concept
The explanation helps me understand how the [software, algorithm, tool] works.	The rank order to the modes of transport, and the use of blocks to indicate the contribution that each factor make to this ordering.
The explanation of how the [software, algorithm, tool] works is satisfying.	We assume that producing an outcome that can evaluate choice of transport could be satisfying (but this requires user testing).
The explanation of the [software, algorithm, tool] is sufficiently detailed.	The output of the recommendation system, in addition to the ranking of choices, is a plan that shows the preferred journey.
The explanation of how the [software, algorithm, tool] works is sufficiently complete.	We believe (despite the opacity of the weighting algorithm) that users will understand how they can modify the ranking through altering the factors.
The explanation is actionable, that is, it helps me know how to use the [software, algorithm, tool]	The outcome is a plan for a journey.

An initial review of the design concept was made against Hoffman et al.'s. (2018) Explainability Checklist (Table 6). This was a useful exercise in the design as it indicated which assumptions we had made in our design and the extent to which these assumptions supported a design that could provide explanation. We made some minor changes to the User Interface on the basis of this review. Our main evaluation involved user testing, as described in the following sections (Table 4).

User trial[3]

Ethical statement

The design of the user trial and processing of data was approved under the ethical procedures of the School of Computer Science, University of Birmingham.

Table 5. Participants' response to the elements in the User Interface.

Element	Response
Factor selection (Figure 4, step 2)	All participants found this straightforward to use; they did not require an explanation as they were able to read the instructions through the interface.
Factors (Figure 4, step 2)	2 participants did not understand the meaning of specific factors such as 'zero emissions'
Ranking (Figure 4, step 3)	3 participants were confused by the information table for re-ranking. 17 participants understood how to rank the factors.
Compare option (Figure 4, steps 4 and 5)	Participants took a while to read this information and understand what it was doing. In general participants understood how the factors and the 'review options' chart correlate. They also understand that the 'ideal recommendations' chart could be compared with the 'review options' chart. 4 participants asked whether they could remove certain routes from the recommendations such as 'car' since they did not own a vehicle.
Chatbot (Figure 4, step 6)	15 participants understood how to use the system. However, 3 participants rushed the reading and when presented with a question within the chatbot they would type in a random answer such as 'okay' or 'Yes' in the incorrect format.

Table 6. Themes from the participant's reflection on the concept map.

Theme	Comment
Recommendation system provides additional factors	Participants would not normally consider factors beyond 'time' and cost'. However, the factors offered by the recommendation system could be added to their list.
Route choice varies on day-to-day basis	Choice of route could depend on time pressures, e.g. appointments, or on the weather. Several participants decided not to rank 'physical health' highly because this would push 'walking' to the top of their recommendation list, resulting in a longer journey time.
Reluctance to add too many factors	An average of three to four features were chosen when using the presented with the 'feature selection' element of the recommender system. Participants were reluctant to add too many features.
Persuasion	Participants felt that the chatbot sought to persuade the user to add factors. A common response was that these factors 'makes sense', or 'why not' or this factor seems like a 'good idea'. In some cases, the recommender system's additional factors challenged the participants original beliefs, e.g. many participants were unaware of the time it would take to cycle from Birmingham City Centre to University of Birmingham; or how the feature of 'mental health' would impact their travelling.
Change in recommended mode of transport	The recommendations changed when speaking with the chatbot, and after inspecting these recommendations; a lot of the answers varied from 'it makes sense', 'this is better/more optimal'.

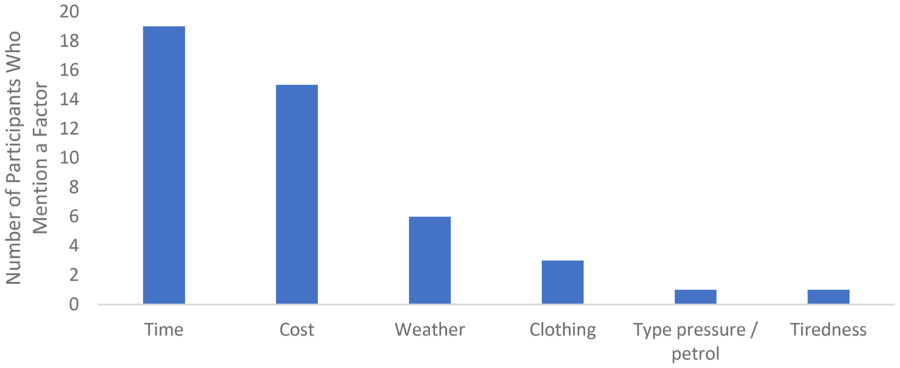


Figure 5. Count of factors mentioned by participants in the initial discussion.

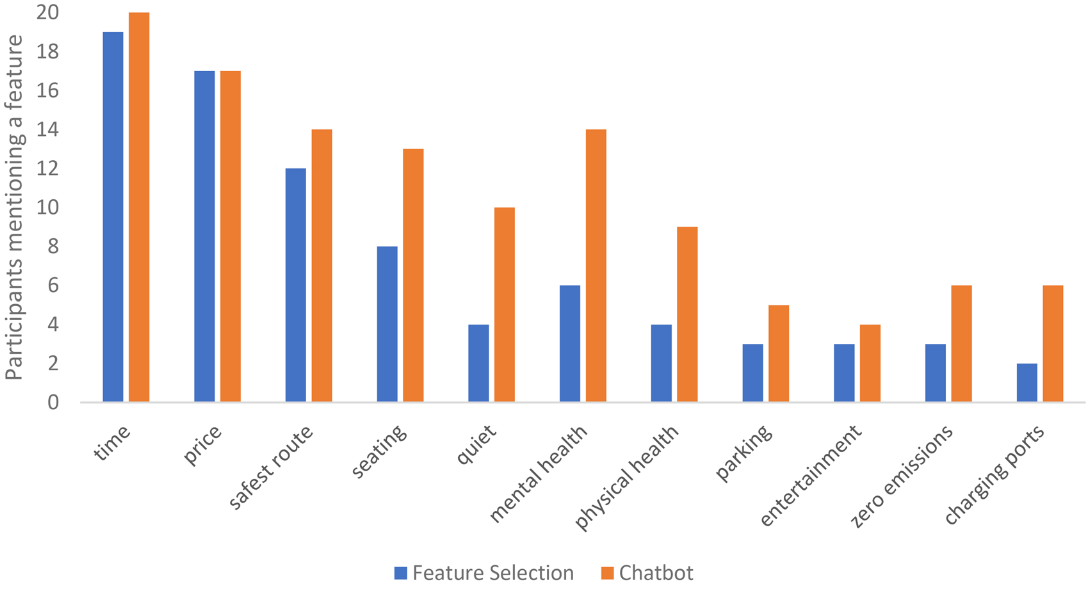


Figure 6. Count of Factors mentioned by participants after interaction. This shows a bar graph with grey (for interactive chart) and blue (for chatbot) of the total occurrence of each factor in user response. This ranges from 20 for time, to less than 5 for entertainment. In all cases, the number of occurrences are higher when participants interact with the chatbot.

Participants

20 participants were involved in this study, participants were either current undergraduate or recent graduates from Universities in the West Midlands of the UK. We did not control for computer experience, but the subjects studied included Computer Science, Engineering,

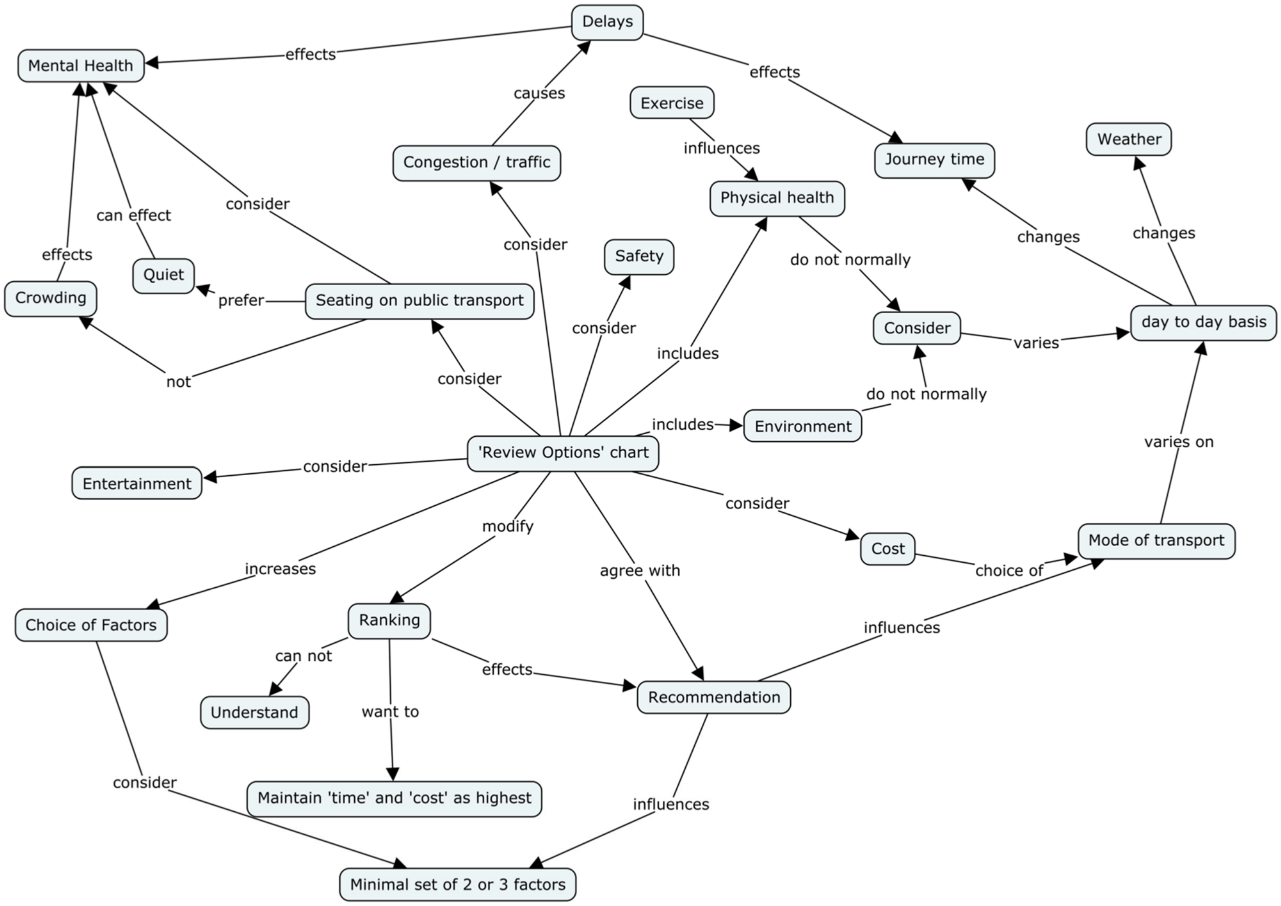


Figure 7. Concept map from participants following interaction with the Recommender System. This shows boxes (labelled with the factors and other concepts mentioned by participants in the Cognitive Walkthrough) connected with arrows to form a concept map.

Economics and English Literature. The age of participants ranged from 21 to 30 years (13 female; 7 male). Participants were resident in Birmingham and all were familiar with the journey from University of Birmingham campus to the City Centre (either as students or as visitors to friends or family).

Procedure

Participants were asked to interact with the Recommender System to define a journey. They were asked, before the interaction began, what factors they normally consider when planning a journey. This involved them answering the question 'what factors do you believe are important when planning a journey from the University campus to the City Centre?' We used this to define the baseline against which we could compare the set of features that were considered following the interaction.

Participants were given an explanation of the use of the Recommender System, using screen-shots that described a journey from Birmingham to Wolverhampton as an example. This demonstrated the ways in which factors were selected and ranked and the use of the chatbot. They were then asked to interact with the chatbot to plan a journey from University to City Centre. As they interacted with the recommender system, we used Cognitive Walkthrough (John and Packer 1995) to encourage them to articulate their impressions of the system's operation. Contemporaneous notes were made during this session. In addition, participants were given prompt questions for each screen of the recommender system, e.g. 'what do think is important on this screen?', 'what do you believe<feature on screen>means?', 'what do you expect to happen when you<perform action>?'.

Analysis

The cognitive walkthrough was subjected to thematic analysis (Bainbridge and Sanderson 1995). Participants were asked about their impressions of the User Interface and the factors that they believe influenced their choice and whether these agreed with those offered by the computer. The factors were summarised in a Concept

Map (Figure 7) produced using C-Map tool.[4] This was constructed during the interviews. Participants were then asked to comment on the Concept Map and these comments were also analysed using thematic analysis.

Results

Participant understanding of the User Interface is summarised in Table 5. While the majority understand these elements and used the recommender system as intended, there was some confusion for 3/20 participants on how the ranking worked.

Figure 5 summarises the factors that participants mentioned prior to interacting with the Recommender System. The most common factors mentioned by participants were 'time' and 'cost/price'. 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 do not 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 that 'weather' could affect their travelling arrangements or the time they leave their house. The mode of transport they would take was mentioned, e.g. some users mentioned 'train' or 'Uber. We note that the factors in Figure 5 are a subset of those elicited from the focus group conducted to define the factors noted above, i.e. {timing, price, emissions, congestion, capacity, number of changes, health, entertainment, charging ports, seating, safety, quiet, parking}

Following the interaction, participants were invited to reconsider their choice of factors. Figure 6 shows the effect of interacting with factor selection (Figure 3, step 2) or chatbot (Figure 3, step 6) on the number of factors mentioned by participants.

Figure 6 indicates that interaction with the Recommender System increases the number of factors considered (and introduces factors additional to those mentioned by the focus group). Interaction with the chatbot increases the number of instances further, particularly those factors relating to the 'ideal' recommendations (Figure 3, step 5). From the Cognitive Walkthrough and choice of factors, we constructed a concept map (Figure 7). The concept map was shared with participants and their responses were analysed into themes. Themes mentioned by at least 5 participants) are shown in Table 6.

Participants preferred their own factors and preferred to focus on a small set (even after seeing the factors offered by the Recommendation System). The majority (17/20) of participants felt that the 'recommendations' derived from physical health 'made sense'. The reasons given for not including these factors in their own decisions 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 factor in their decisions. 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, or because they did not own a car. The chatbot helped participants understand why factors were chosen, and once they understood the reasoning for this factor, they could then re-rank and alter this within their charts to receive a recommendation more closely related to their preferences, e.g. some participants did not understand how safety would lead to a higher rating for walking and thought the car was safer (despite differences in accident statistics that the chatbot's database included).

While participants were able to reflect on differences between the feature sets, it is a moot point as to whether either participant or computer has produced the 'best' set of features. Users were able to make sense of the computer's feature sets and how these corresponded to a specific recommendation. However, the question of how we might reconcile disagreements (especially in application domains that have safety implications or where a correct judgement is required) is beyond the scope of our study. We appreciate that there will be situations in which human or computer might make mistakes in their final decision. Our aim is to surface the features that have been selected in support of the choice; to allow the human to reflect on their own preferences as expressed in their choice of features, and to allow the human to make sense of the computer's choice. Where is a discrepancy between choices this can be explored through the chatbot to determine why specific features are relevant to the choice being proposed.

Participants were able to ask questions of the chatbot to gain further insight into the computer's choice and the features that contributed to this. In addition, participants were able to explore their own selection of features through the Feature Selection screen, where they could edit their set of features or change the feature weights. In this way, participants had the opportunity to explore alternative choices for their journey.

In terms of presenting the recommendations in a manner suitable to participants, we used a familiar context and sought to present alternative perspectives on this. Thus, all participants were familiar with making the journey from the University to the City Centre and we wanted to see if they were able to see how making the journey in different ways could be beneficial for their health or the environment. All participants accepted that there were alternatives, although few were willing to change the own choice of journey. This, for our project, is less important than demonstrating that we could present the computer output in a way that participants could understand. In other words, our aim was to demonstrate our concept of explanation rather than to force behaviour change.

Discussion

A framework is developed to highlight this concept, and this is instantiated to show how different types of explanation can occur; each of which requires different means of support. Primarily, an explanation involves agreement on the features (in data sets or a situation) to which explainer and explainee attend, and agreement as to why these features are relevant (and we propose three levels of relevance, i.e. 'cluster' in which a group of features will typically occur together; 'belief' which defines a reason as to why such a cluster will occur; 'policy' which justifies the belief and relates this to action). Relating our work to ongoing research in XAI, we have proposed a formal approach which we believe elaborates on prior work, such as Holzinger, Carrington, and Müller (2020) and Rosenfeld and Richardson (2019), by considering the human as an active partner in an explanatory discourse. In our approach, the purpose of explanation is to ensure that explainer and explainee reach agreement of the 'features' that are being used to support the explanation, and on the 'relevance' of these features. We believe that AI-centric approaches tend to only present the features used by the AI system and do not allow users to either define their own features or to challenge those used by the AI system. In our approach, users are invited to provide and rank features that reflect their preferences. This has the added benefit of encouraging users to reflect on their preferences. We allow users to compare the weighted features of their preferences with those of a computer, with the opportunity to either modify their own preferences or to engage, through a chatbot, in conversation to discover why the computer has weighted the features as it has. This allows the user to appreciate how the computer has defined relevance through its choice of weighted features. We believe that the framework we propose is sufficient to provide the basis for future XAI developments.

The evaluation indicates that there could be a disagreement between the factors that participants expressed and those offered by the recommendations towards specific decisions (i.e. zero emissions or physical health). Agreement (on features and on relevance) depends on the knowledge and experience of explainer and explainee, and much of the process of explanation involves ensuring alignment in terms of knowledge and experience. Thus, 'Explanation' is the process by which common ground is established and maintained. We propose that the process through which the prototype Recommender System can support explanatory dialogues around aligning or challenging the definition of situation or relevance can enhance the development of common ground. In this respect, while of very limited functionality, the Recommender System fulfilled its purpose by encouraging participants to think about factors beyond the ones that they had initially stated, and by helping participants to appreciate why the Recommender System was proposing its 'ideal' recommendations. This does not mean that either the Recommender System or user have provided 'correct' answers. Future work could explore the relationship between the accuracy of recommendations and user preferences,

Table 7. Guidelines on developing XAI to support explanatory discourse.

Guideline	Rationale	Evidence from our User Trial
Explanations should include relevant *Features*	Explainer and Explainee should agree key features of the situation.	Users could select factors that they deemed important in the choice of journey. These were contrasted with factors that the computer used for a specific recommendation. Users were able to discern and reflect on any differences between the factors.
Explanations should highlight *Relevance*	The relationship between features of a situation and the event being explained should be plausible in terms of a concept of Relevance agreed between Explainer and Explainee.	Users were able, through a chatbot, to discuss the factors and how these contributed to the computer's recommendation. Further development of the chatbot will allow the human to ask 'what-if' (i.e. counter-factual questions), although in the current version this can be explored by the user selecting different features or changing their weights.
Explanations should be Framed to suit the *audience*	The explanation should align with the explainee's understanding of the situation and their goals.	We focussed on the specific task of making a familiar journey and explored ways in which choices for the journey could be affected by concerns for personal health or the environment. Further work could explore unfamiliar or new journeys.
Explanations should be (where appropriate) *actionable*	The explainee should be given information that can be used to perform actions and behaviours.	In addition to suggesting a means of making a journey, the computer provides detailed guidance on how the achieve this (i.e. with a route map and comments on how to make the journey).

but our focus in this paper was to explore whether our process model of explanation offers support to developing explainable user interfaces.

We have demonstrated how a design for a recommender system can be developed from our explanation 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 human-centred explainable AI. From our process model, design exercise, and user trial, we offer guidelines to support explanatory discourse (Table 7).

Notes

1. *AI in the UK: Ready, Willing and Able?*, report, UK Parliament (House of Lords) Artificial Intelligence Committee, 16 April 2017, paragraph 12; https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/10002.html.
2. https://www.bbc.co.uk/programmes/m001216k.
3. The user trial was conducted as part of an MSc dissertation project by PK under the supervision of CB.
4. https://cmap.ihmc.us.
5. Example created from Radečić, 2020, http://www.towardsdatascience.com/lime-how-to-interpret-machine-learning-models-with-python-94b0e7e4432e, using the Kaggle 'Wine Quality' dataset.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

Part of the work reported in this paper was supported by a grant from the ESRC/CREST [Project: 27, 2020-2021].

References

Adadi, A., and M. Berrada. 2018. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)." *IEEE Access* 6: 52138–52160. doi:10.1109/ACCESS.2018.2870052.

Alufaisan, Y., L. R. Marusich, J. Z. Bakdash, Y. Zhou, and M. Kantarcioglu. 2021. "Does Explainable Artificial Intelligence Improve Human Decision-Making?" *Proceedings of the AAAI Conference on Artificial Intelligence*, 6618–6626. Washington, DC: Association for the Advancement of Artificial Intelligence.

Barredo Arrieta, Alejandro, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI." *Information Fusion* 58: 82–115. doi:10.1016/j.inffus.2019.12.012.

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

Baber, C., E. McCormick, and I. Apperley. 2021. "A Human-Centered Process Model for Explainable AI." *Naturalistic Decision Making and Resilience Engineering Symposium 2021*.

Bainbridge, L., and P. Sanderson. 1995. "Verbal Protocol Analysis." In *Evaluation of Human Work: A Practical Ergonomics Methodology*, edited by J. R. Wilson and E. N. Corlett, 2nd ed., 169–201. London: Taylor and Francis.

Bansal, G., B. Nushi, E. Kamar, E. Horvitz, and D. S. Weld. 2021. "Is the Most Accurate AI the Best Teammate? Optimizing AI for Teamwork." *Proceedings of the AAAI Conference on Artificial Intelligence* 35 (13): 11405–11414. doi:10.1609/aaai.v35i13.17359.

Caraban, A., E. Karapanos, D. Gonçalves, and P. Campos. 2019. *23 Ways to Nudge: A Review of Technology-Mediated Nudging in Human-Computer Interaction, CHI'19*, 1–15. New York: ACM.

Carton, S., Q. Mei, and P. Resnick. 2020. "Feature-Based Explanations Do Not Help People Detect Misclassifications of Online Toxicity." *Proceedings of the International AAAI Conference on Web and Social Media* 14: 95–106. doi:10.1609/icwsm.v14i1.7282.

Clark, H. H. 1991. *Using Language*. Cambridge: Cambridge University Press.

Clark, H. H., and S. E. Brennan. 1991. "Grounding in Communication." In *Perspectives on Socially Shared Cognition*, edited by L. B. Resnick, J. Levine, and S. D. Teasley, 127–149. Washington, DC: American Psychological Association.

De Graaf, M. M., and B. F. Malle. 2017. "How People Explain Action (and Autonomous Intelligent Systems Should Too)." *2017 AAAI Fall Symposium Series*. Washington, DC: Association for the Advancement of Artificial Intelligence.

Edwards, L., and M. Veale. 2017. "Slave to the Algorithm: Why a Right to an Explanation is Probably Not the Remedy You Are Looking for." *Duke Law & Technology Review* 16: 18.

Ehrlich, K., S. E. Kirk, J. Patterson, J. C. Rasmussen, S. I. Ross, and D. M. Gruen. 2011. "Taking Advice from Intelligent Systems: The Double-Edged Sword of Explanations." In *Proceedings of the 16th International Conference on Intelligent User Interfaces*, 125–134. New York: Association for Computing Machinery.

Guidotti, R., A. Monreale, F. Giannotti, D. Pedreschi, S. Ruggieri, and F. Turini. 2019. "Factual and Counterfactual Explanations for Black Box Decision Making." *IEEE Intelligent Systems* 34 (6): 14–23. doi:10.1109/MIS.2019.2957223.

Hempel, C. G. 1924. "The Function of General Laws in History." *The Journal of Philosophy* 39 (2): 35–48. doi:10.2307/2017635.

Hoffman, R. R., G. Klein, and S. T. Mueller. 2018. "Explaining Explanation for "Explainable AI." In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *62*, 197–201. Los Angeles, CA: Sage. doi:10.1177/1541931218621047.

Hoffman, R. R., S. T. Mueller, G. Klein, and J. Litman. 2018. "Metrics for Explainable AI: Challenges and Prospects." arXiv:1812.04608.

Holzinger, A., A. Carrington, and H. Müller. 2020. "Measuring the Quality of Explanations: The Systems Causability Scale (SCS)." *Kunstliche Intelligenz* 34 (2): 193–198.) doi:10.1007/s13218-020-00636-z.

John, B. E, and H. Packer. 1995. *Learning and Using the Cognitive Walkthrough Method: A Case Study Approach, CHI'95*, 429–436. New York: ACM.

Klein, G., R. Hoffman, S. Mueller, and E. Newsome. 2021. “Modeling the Process by Which People Try to Explain Complex Things to Others.” *Journal of Cognitive Engineering and Decision Making* 15 (4): 213–232. doi:10.1177/15553434211045154.

Klein, G., J. K. Phillips, E. L. Rall, and D. A. Peluso. 2007. “A Data–Frame Theory of Sensemaking.” In *Expertise Out of Context*, 118–160. Psychology Press; ILO.

Krishna, S., T. Han, A. Gu, J. Pombra, S. Jabbari, S. Wu, and H. Lakkaraju. 2022. “The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective.” arXiv Preprint arXiv:2202.01602

Langley, P. 2019. “Varieties of Explainable Agency.” In *International Conference on Automated Planning and Scheduling Workshop on Explainable Planning*. CA: ICAPS.

Leddo, J., R. P. Abelson, and P. H. Gross. 1984. “Conjunctive Explanations: When Two Reasons Are Better than One.” *Journal of Personality and Social Psychology* 47 (5): 933–943. doi:10.1037/0022-3514.47.5.933.

Maathuis, C. 2023. “Human Centered Explainable AI Framework for Military Cyber Operations.” In *IEEE Military Communications Conference (MILCOM)*, 260–267. New York: Institute of Electronic and Electrical Engineers.

Maynard Smith, J., and D. Harper. 2003. *Animal Signals*. Oxford: Oxford University Press.

McClure, J., D. Hilton, J. Cowan, L. Ishida, and M. Wilson. 2001. “When Rich or Poor People Buy Expensive Objects: Is the Question How or Why.” *Journal of Language and Social Psychology* 20 (3): 339–357. doi:10.1177/0261927X01020003004.

Miller, T. 2019. “Explanation in Artificial Intelligence: Insights from the Social Sciences.” *Artificial Intelligence* 267: 1–38. doi:10.1016/j.artint.2018.07.007.

Mueller, S. T., R. R. Hoffman, W. Clancey, A. Emrey, and G. Klein. 2019. “Explanation in human-AI Systems: A Literature Meta-Review, Synopsis of Key Ideas and Publications, and Bibliography for Explainable AI.” arXiv:1902.01876

Mueller, S. T., E. S. Veinott, R. R. Hoffman, G. Klein, L. Alam, T. Mamun, and W. J. Clancey. 2021. “Principles of Explanation in Human-AI Systems.” arXiv:2102.04972/AAAI'21 – explainable agency in artificial intelligence workshop

Neerincx, M. A., J. van der Waa, F. Kaptein, and J. van Diggelen. 2018. “Using Perceptual and Cognitive Explanations for Enhanced Human-Agent Team Performance.” In *EPCE 2018, LNCS (LNAI), 10906*, edited by D. Harris, 204–214. Cham: Springer

Ribeiro, M. T., S. Singh, and C. Guestrin. 2016. “Why Should i Trust You?” Explaining the Predictions of Any Classifier.” In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. New York: Association for Computing Machinery.

Rosenfeld, A., and A. Richardson. 2019. “Explainability in Human–Agent Systems.” *Autonomous Agents and Multi-Agent Systems* 33 (6): 673–705. doi:10.1007/s10458-019-09408-y.

Salmon, W. 1998. *Causality and Explanation*. Oxford: Oxford University Press.

Schwalbe, G., and B. Finzel. 2023. “A Comprehensive Taxonomy for Explainable Artificial Intelligence: A Systematic Survey of Surveys on Methods and Concepts.” *Data Mining and Knowledge Discovery*. doi:10.1007/s10618-022-00867-8.

Springer, A. 2019. “Enabling Effective Transparency: Towards User-Centric Intelligent Systems.” In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 543–544. Washington, DC: Association for the Advancement of Artificial Intelligence.

Stanton, N A., R. Stewart, D. Harris, R J. Houghton, C. Baber, R. McMaster, P. Salmon, G. Hoyle, G. Walker, M S. Young, M. Linsell, R. Dymott, and D. Green. 2006. “Distributed Situation Awareness in Dynamic Systems: Theoretical Development and Application of an Ergonomics Methodology.” *Ergonomics* 49 (12–13): 1288–1311. doi:10.1080/00140130600612762.

Tversky, A., and D. Kahneman. 1983. “Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment.” *Psychological Review* 90 (4): 293–315. doi:10.1037/0033-295X.90.4.293.

Wang, X., and M. Yin. 2021. “Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making.” In *26th International Conference on Intelligent User Interfaces*, 318–328. New York: Association for Computing Machinery. doi:10.1145/3397481.3450650.

Wilkinson, D., Ö. Alkan, Q. Liao, M. Mattetti, I. Vejsbjerg, B. Knijnenburg, and E. Daly. 2021. “Why or Why Not? The Effect of Justification Styles on Chatbot Recommendations.” *ACM Transactions on Information Systems* 39 (4): 1–21. doi:10.1145/3441715.