Appendix:
Explanation Forms, Design Support, Study Method, Quantitative Results, and Detailed Qualitative Results

In this Appendix, we provide details on the user study method, materials, and results. Specifically, we present the figures of visual representations for each explanation form in §1. The human-centered design methodology support is in §2. In §3, we give further details on the user study design and data analysis methods. §4 reports the quantitative results on participants' selection and ranking of the explanation forms. §5 provides detailed qualitative results as supplementary material for the main reporting in the manuscript, for the reader to check more context on users' interpretation and requirements for each explanation form. The participants' demographic information is in §6. We further attach the study material in the user study at the end in §7, including the demographic questionnaire and the interview material of the task, explanation goals, and explanation forms.

Contents

1 Visual Representations of the Explanation Forms

1.1 Feature-Based Explanation

1.1.1 Feature Attribution

Visual representation: Its visual representations largely depend on the feature data type. For image and text data, overlaying a **saliency map** or color map on the input is the most common visualization. It uses sequential colors to code the fine-grained feature importance score for each individual feature (could be a pixel for image input, a word for text data). For image/video input data, other popular visualizations include using *segmentation masks* or *bounding boxes* on important image objects/parts.

To visualize multiple feature attributions for tabular or text data, a **bar chart** is a typical choice. Variations of the bar chart include waterfall plot, treemap, wrapped bars, packed bars, piled bars, Zvinca plots, and tornado plot. Compared to a bar chart that shows a point estimation of feature importance, a *box plot* can be used to visualize the probabilistic distribution of the feature importance score. Its variations include violin plot and swarm plot that show more detailed data distribution and skewness.

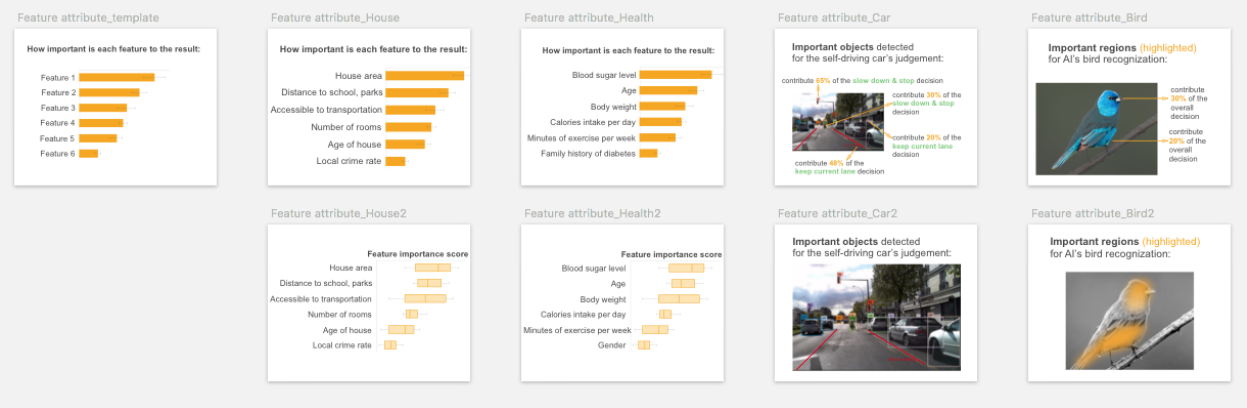


Figure 1: Visual representations of feature attribution. The first column is the template with feature names shown by placeholders. The following columns are the visual representation cards used in the four tasks (House, Health, Car, Bird) in the user study. Different visual variations are shown in the second row.

1.1.2 Feature Shape

Visual representation: For a continuous feature (such as height, temperature, i.e.: measurement on a scale), a **line chart** is the most common visualization, depicting whether the relationship between the feature and outcome is monotonic, linear, or more complex. The line chart can be accompanied by a scatter plot detailing the position of individual data points.

For a categorical feature (such as gender, season), a *bar chart* can be used.

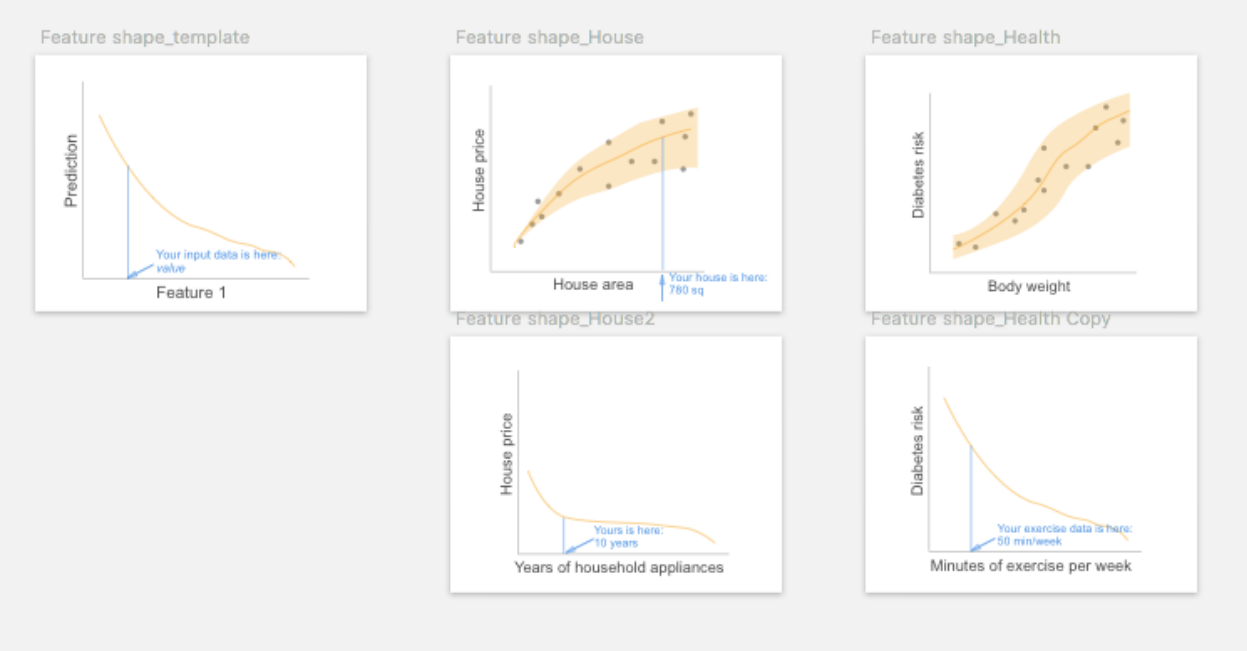


Figure 2: Visual representations of feature shape. The first column is the template with feature names shown by placeholders. The following columns are the visual representation cards used in the two tasks (House, Health) in the user study. Different visual variations are shown in the second row.

1.1.3 Feature Interaction

Visual representation: **2D or 3D heatmap** is usually used to visualize the combined effect of feature interactions on prediction. Limited by the visualization, a heatmap can show feature interactions for at most three features (using 3D heatmap). More complicated multiple paired feature-feature interactions can be visualized using matrix heatmap, node-link network, or contingency wheel.

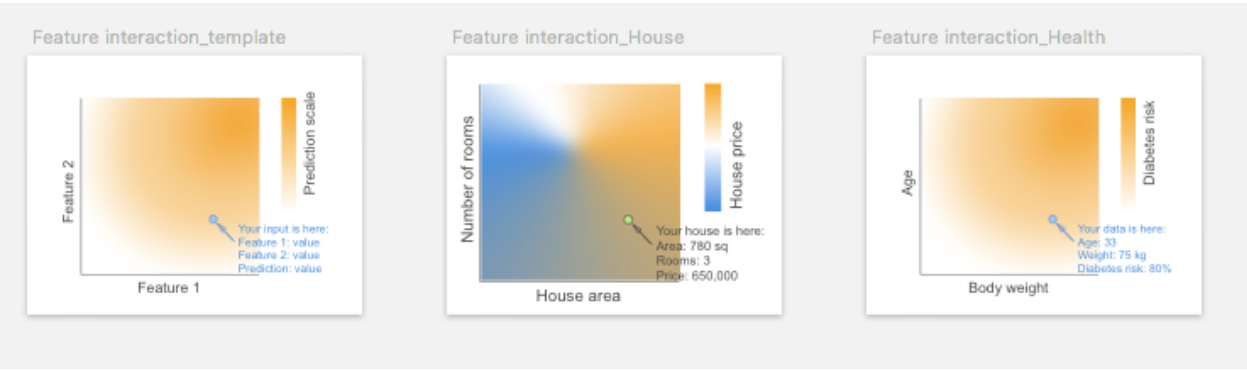


Figure 3: Visual representations of feature interaction. The first column is the template with feature names shown by placeholders. The following columns are the visual representation cards used in the two tasks (House, Health) in the user study.

1.2 Example-Based Explanation

1.2.1 Similar Example

The differences among similar, typical, and counterfactual examples are listed in Table A1: For a similar example, although it shares *similar features* with the query instance, their predictions may be *the same or different*. Whereas for a counterfactual example, it not only shares *similar features* with the query instance, but also has a *different prediction*. For a typical example, it has *the same prediction* as the query instance, regardless of their features.

Explanation Form	**Features**	**Prediction**
Similar Example	similar	the same or different
Typical Example	similar or different	the same
Counterfactual Example	similar	different

Table A1: **Distinctions among the three example-based explanation forms** by comparing their features and prediction with the query instance.

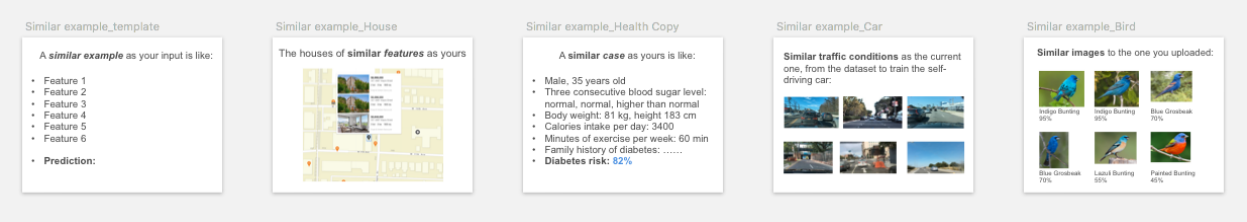


Figure 4: Visual representations of similar example. The first column is the template with feature names shown by placeholders. The following columns are the visual representation cards used in the four tasks (House, Health, Car, Bird) in the user study.

1.2.2 Typical Example

Visual representation: For similar and typical examples, it is straightforward to show several examples with their corresponding predictions.

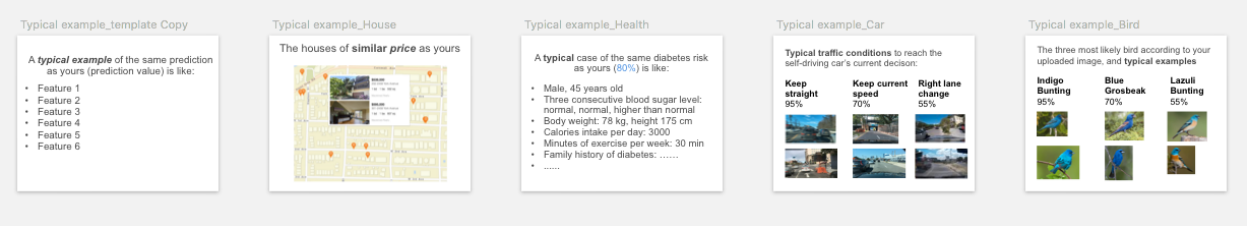


Figure 5: Visual representations of typical example. The first column is the template with feature names shown by placeholders. The following columns are the visual representation cards used in the four tasks (House, Health, Car, Bird) in the user study.

1.2.3 Counterfactual Example

We noted that counterfactual explanations can also be expressed as counterfactual features or rules. However, a counterfactual feature/rule can not be a standalone explanation in an XAI system, they must reside within a certain context by assuming all other features are constant. To make the explanation information complete, we include the counterfactual explanation in the form of example.

Visual representation: Counterfactual examples can be shown as two instances and their predictions, with their **counterfactual/contrasting features highlighted**, or a **transition** from one instance to the other by gradually changing the counterfactual features.

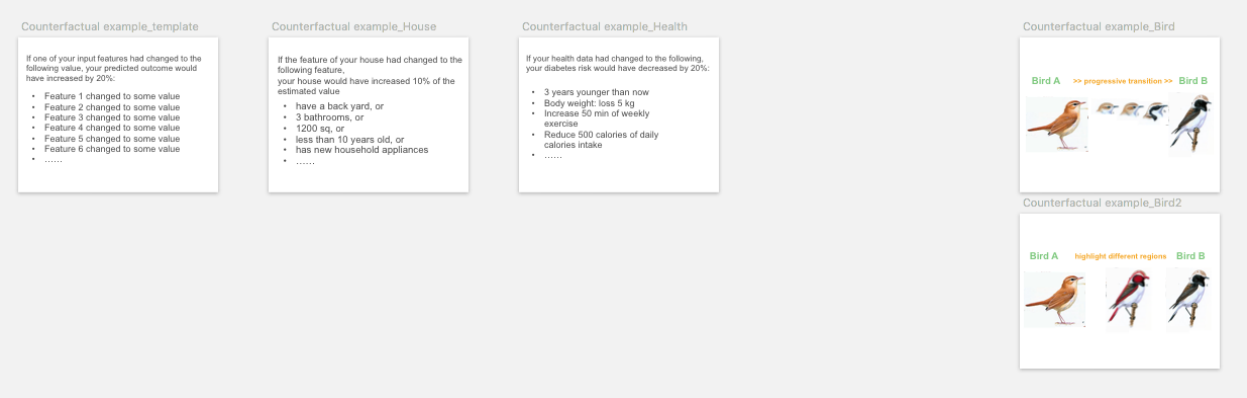


Figure 6: Visual representations of counterfactual example. The first column is the template with feature names shown by placeholders. The following columns are the visual representation cards used in the three tasks (House, Health, Bird) in the user study. We did not include a counterfactual example for the Car task, therefore we leave the column black. Different visual variations are shown in the second row.

1.3 Rule-Based Explanation

We note decision rule and decision tree actually carry similar explanation information. However, since they are usually generated by different XAI algorithms, and their representation formats (text vs. diagram) are different to users, we included them as two separate explanation forms.

1.3.1 Decision Rule

Visual representation: Decision rules are usually represented using **text**. Other representing formats include table [6] or matrix [17] to align, read, and compare rule clauses more easily.

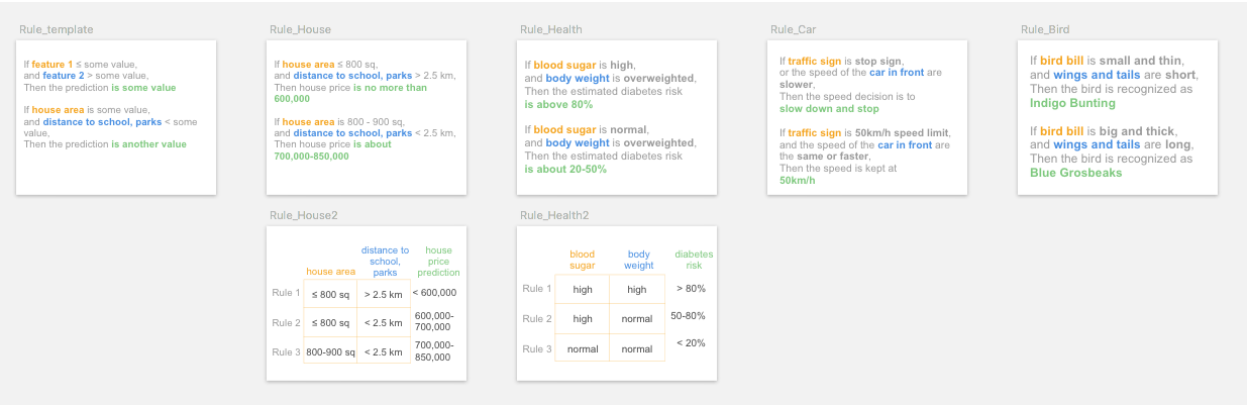


Figure 7: Visual representations of decision rule. The first column is the template with feature names shown by placeholders. The following columns are the visual representation cards used in the four tasks (House, Health, Car, Bird) in the user study. Different visual variations are shown in the second row.

1.3.2 Decision Tree

Visual representation: The most common representation is to use a node-link **tree diagram**. Other visual representations to show the hierarchical structure include treemap, cladogram, hyperbolic tree, dendrogram, and flow chart.

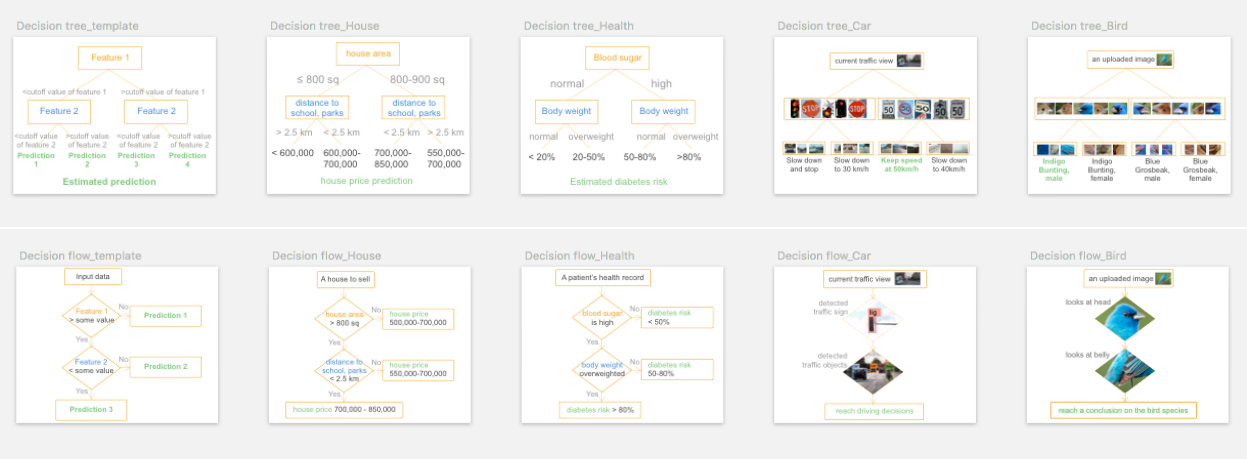


Figure 8: Visual representations of decision tree (1st row) and decision flow (2nd row). The first column is the template with feature names shown by placeholders. The following columns are the visual representation cards used in the four tasks (House, Health, Car, Bird) in the user study.

1.4 Contextual information

To provide necessary context and background for a more complete explanation, we additionally include contextual information in the end user-friendly explanation forms, include:

Input x.

Output y.

Performance: Model's performance metrics (such as prediction accuracy, confusion matrix, ROC, mean squared error) help end users to judge a model's overall decision quality and set a proper expectation on model's capability.

Dataset: It is the proper description of the training and validation dataset, such as data distribution, and how the data were collected.

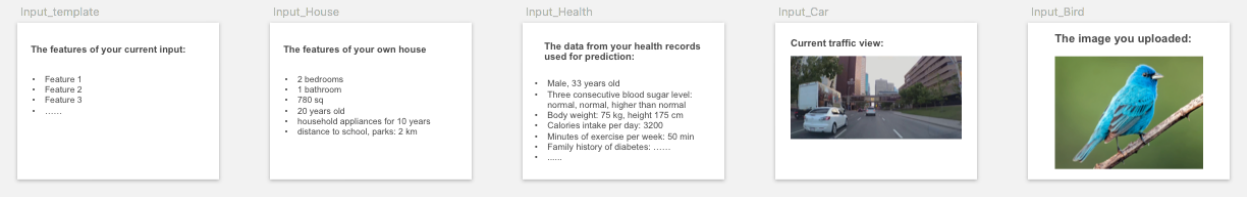


Figure 9: Visual representations of input. The first column is the template with feature names shown by placeholders. The following columns are the visual representation cards used in the four tasks (House, Health, Car, Bird) in the user study.

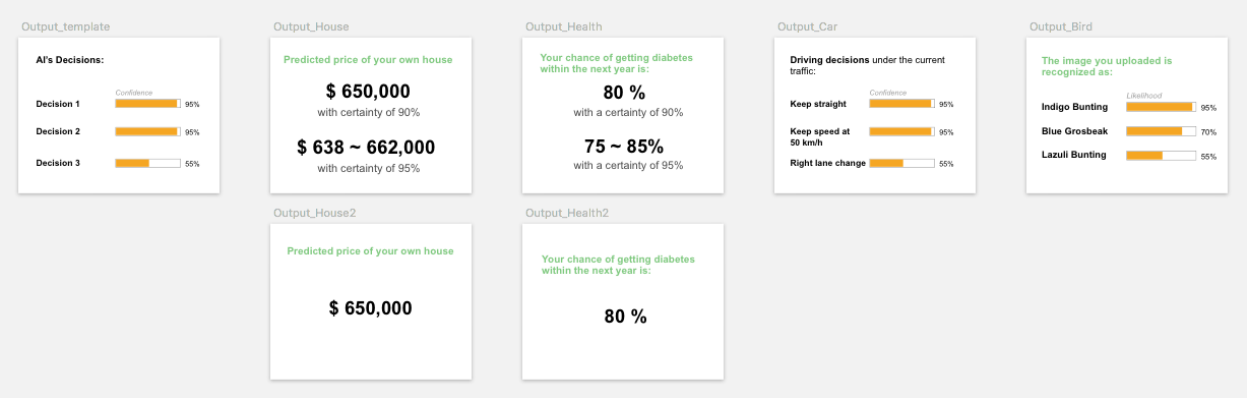


Figure 10: Visual representations of output. The first column is the template shown using placeholders. The following columns are the visual representation cards used in the four tasks (House, Health, Car, Bird) in the user study. Different visual variations are shown in the second row.

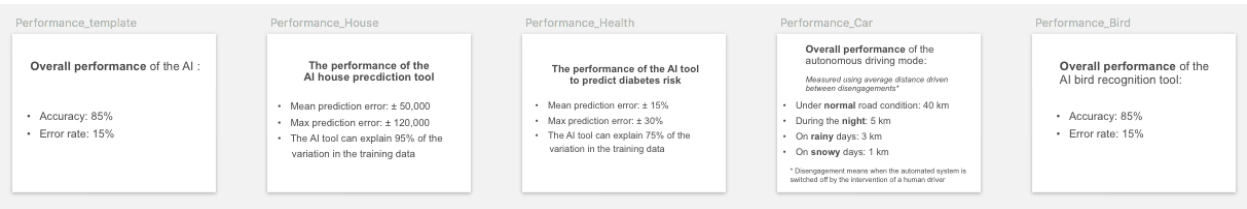


Figure 11: Visual representations of performance. The first column is the template. The following columns are the visual representation cards used in the four tasks (House, Health, Car, Bird) in the user study.

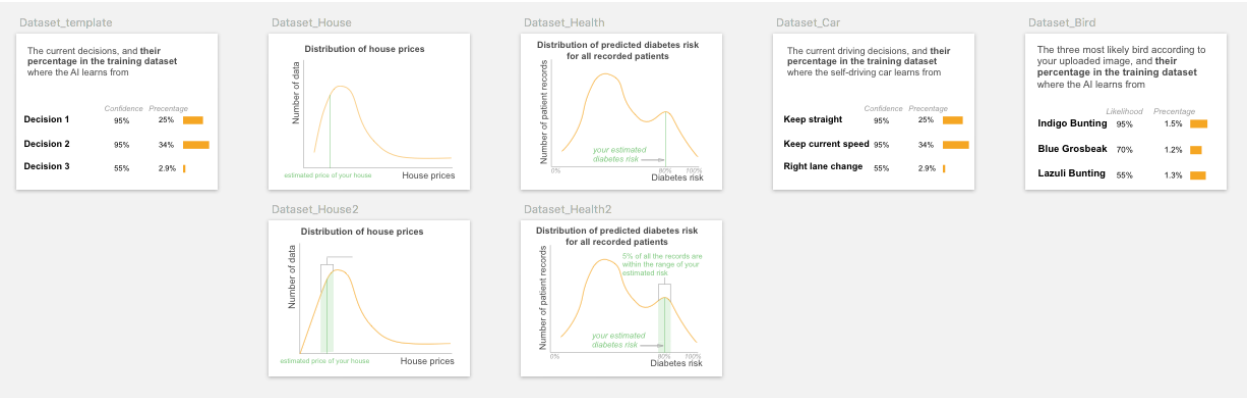


Figure 12: Visual representations of dataset. The first column is the template shown using placeholders. The following columns are the visual representation cards used in the four tasks (House, Health, Car, Bird) in the user study. Different visual variations are shown in the second row.

2 End-User-Oriented Design Support

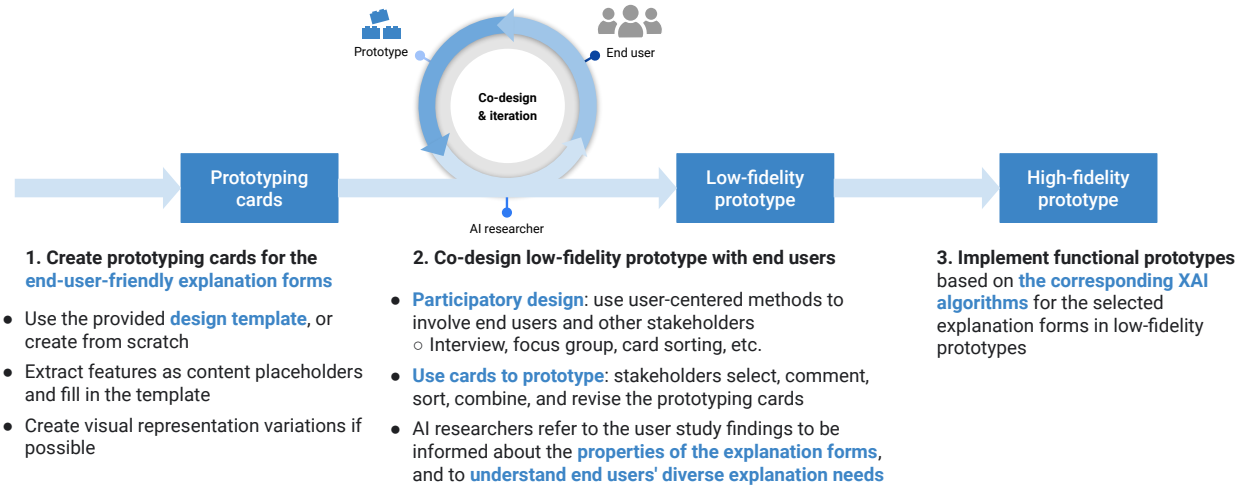


Figure 13: **The suggested prototyping workflow with end users to gain their insights on the requirements for XAI techniques.** Step 2 is an iterative process with end users. The blue highlights are design support and materials provided by this work.

We provide a human-centered design methodological support on the co-design and prototyping methods with end users. It facilitates AI researchers to communicate with end users to understand their requirements for XAI techniques. The following design steps (Fig. 13) basically follow the user study procedure based on the explanation forms.

1. **Create prototyping cards from explanation forms**

   The AI researcher starts by **manually extracting several interpretable features** given the AI task and input/output data type. For example, for tabular data, the features could be the column names that describing the input instance, such as house size, age, and location. For image data, the features could be saliency image part or object for recognition, such as cars, traffic signs, or pathological appearance of a disease on chest X-ray. As quick prototyping, the feature content may not necessarily reflect the real content generated by XAI algorithms. They serve as content placeholders to be filled in the prototyping card templates.

   Then the AI researcher can use the prototyping card template from the user study, and **fill in the template with the above extracted features**. The design templates visualize the explanation forms as user interface elements. Designers can also create their own templates from scratch by referring to the design examples.

   To encourage brainstorming and divergent thinking during the conversation with end users, the AI researcher may **prepare multiple variations** for some particular explanation forms, by varying its visual representation (e.g.: graphics or text) and interface layout, alternating contents from brief to details, and providing different options, such as whether to use pre-defined or user-defined contrastive outcome on counterfactual example, whether to give users the option to set a threshold level for feature attribution. Each explanation form and its variations are presented on individual prototyping cards.

2. **Co-design and iterate low-fidelity prototype with end users**

   With the prepared prototyping cards, the AI researcher then can meet and discuss with the target end users and/or other stakeholders of the XAI system, and apply user-centered methods (informally or formally), such as: interview, focus group, and card sorting. The communication aims to use the created cards as a prototyping tool to understand users' needs and requirements for their explanation goals, and incorporate end users' perspectives in XAI technique co-design, iteration, and evaluation process.

To quickly create a low-fidelity prototype from the prototyping cards (prepared in Step 1), the **end users can select, rank, combine, modify the prototyping cards**, or sketch new ones. In this process, the AI researcher may ask users why they selected or did not select a card, and their rationals for making such a combination, whether the combination could fulfill their requirements, and what is lacking in the current prototype. Users can also comment on and revise each variation of the same explanation form. With the tangible prototyping cards, the AI researcher can know in-details about user's specific requirements on the XAI technique.

After the initial communication with users, AI researchers need to synthesize users comments and decide one or several prototype designs (such as using majority voting). Then based on the prototyping card ranking and combination, the AI researcher may create low-fidelity prototypes, and continue to seek user and/or other stakeholders' feedback and iterate the prototypes.

During the above process, the AI researcher may refer to the user study findings to be informed about the properties of the explanation forms (pros, cons, applicable explanation goals, and design implications in Section 5), and to understand end users' diverse explanation goals (to calibrate trust, detect bias, resolve disagreement with AI, etc.)

3. **Build a functional prototype**

   After the above co-design process and several rounds of iteration, the AI researcher may have a rough idea on users' requirement, and may start to build the XAI techniques as a high-fidelity prototype for further test and evaluation with end users.

3 User Study Method

We conducted a user study that utilized interview, card selection & sorting, and prototyping methodologies. It aims to elicit users' interpretation on the pros, cons, applicable explanation goals, and their requirements for XAI algorithms manifested as explanation forms.

3.1 Participants and Recruitment

We recruited layperson participants via a convenience sampling method by advertising posters at public libraries, community centers, and online community boards in the anonymous area over a 3-month period in 2019. The inclusion criteria were: 1) adult (19 years old and above); and 2) do *not* have prior technical knowledge in machine learning, data science, or artificial intelligence. A total of 32 participants were enrolled in the study (Female = 16; Age: 38.2±16.0, range 19-73). Participants' occupations covered a variety of industries e.g.: technology, design, car insurance, finance, psychology, construction, sales, food/cooking, law, healthcare, government/social services, and retired. For participants who use AI in work or life (6 participants, 19%), they used AI software such as Google Assistant to play music, navigate traffic, chat with clients, and help drive investment decisions. Fig. 14 shows the distribution of participants' age, educational background, familiarity with and attitudes towards AI. Participants' detailed demographics are in Table 6. The participants were thanked with $25 CAD for their time and effort in the study.

3.2 The Interview Instrument

3.2.1 Critical Decision-Making Tasks

We focus the scope of the study on AI-assisted critical decision-support tasks, where explanations have high utility as shown in previous research [5, 11, 7], and AI could not be delegated to have full automation because of the high-stakes nature of the tasks and the liability issue. In this study phase, we did not included domain experts. Therefore, we deliberately designed the tasks so that decisions can be made based on common sense without requiring domain knowledge. We designed four decision-making tasks reflecting the diversity of AI-supported critical decision-making. They are:

1. **House** task: users use AI to get a proper estimate of their house price.

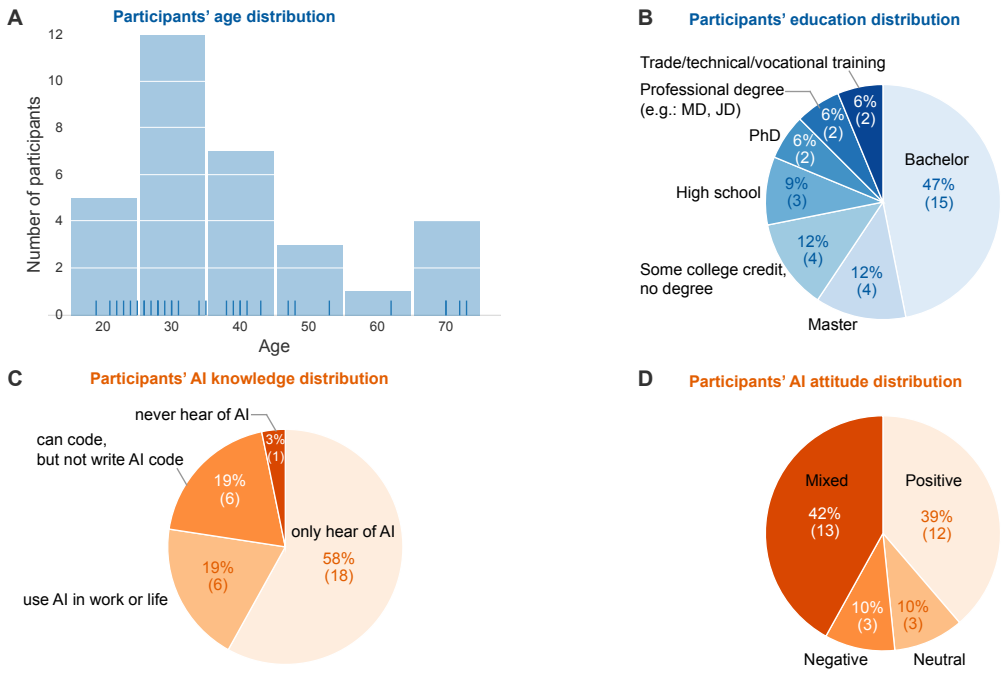


Figure 14: **Participants' demographic information**.
(**A**) Histogram of participants' age distribution. The sticks on the horizontal axis show each participant's age.
(**B**) Pie chart on participants' educational level. Numbers in parentheses represent the number of participants in that category.
(**C**) Pie chart on participants' familiarity with AI.
(**D**) Pie chart on participants' attitudes towards AI; Positive attitudes include *"interested "* in and *"excited"* to use AI; Negative attitudes consist of *"skeptical"* and *"concerned"* about AI; A mixed attitude means participants hold both positive and negative attitudes towards AI.

2. **Health** task: users use AI to predict his/her diabetes risk.
3. **Car** task: users decide whether to buy an autonomous driving vehicle.
4. **Bird** task: users use AI bird recognition tool to prepare for an important biology exam.

The four tasks are critical decision-making scenarios, because their decisions have significant consequences on one's health and life (Health and Car Task), finance (House Task), or education (Bird Task). The four tasks covered common input data types of tabular, sequential, image and video data. The corresponding datasets of the four tasks are publicly available (Table A2), so that the resultant low-fidelity prototypes from this user study can be actualized as high-fidelity functional prototypes for task-specific studies in future work.

AI-Assisted Critical Decision-Making Tasks		**HOUSE TASK** **Sell a house**	**HEALTH TASK** **Check diabetes risk**	**CAR TASK** **Buy a self-driving car**	**BIRD TASK** **Prepare for exam**
					
Explanation Goals	**Calibrate Trust**	You doubt whether to trust the AI tool or not	You doubt whether to trust the software prediction on your diabetes risk	n/a	You don't know whether to trust the results from the website or not
	Ensure Safety	n/a	n/a	You need to know whether the autopilot mode is safe and reliable	n/a
	Detect Bias	n/a	You doubt whether the software will perform the same among people with different gender, age, or ethnicity group	You want to know if the autopilot mode performs robustly under varying road, weather, and light conditions.	n/a
	Unexpected Prediction: Disagreement with AI	AI's prediction aligns/does not align with your own estimation	You maintain good health with no major diseases or a family history of diabetes/Diabetes tends to run in your family, and you're afraid of getting it someday, and AI predicts your chance of getting diabetes is low/high	You notice the car sometimes drives much slower than the expected speed limit	The results sometimes do not align with your knowledge
	Differentiation Similar Instances	n/a	n/a	n/a	In the exam, you need to write a short statement to differentiate different birds
	Learn from AI	n/a	n/a	n/a	Is it a good tool to improve your learning and help you know more about bird taxonomy?
	Improve the Predicted Outcome	You need to decide whether to do a renovation or replacement of appliances to increase your house value, and which action is the most cost-effective	You want to know how to adjust your lifestyle accordingly to lower the risk of diabetes	n/a	n/a
	Communicate with Stakeholders	You need to communicate your decision with your family	You need to need to inform family members and consult your doctor	You need to communicate with your family about your judgment on the car's safety	n/a
	Generate Reports	n/a	n/a	n/a	In the exam, you need to write a short statement on how you recognize the bird as such a species
	Multi-Objectives Trade-Off	n/a	You're aware that the insurance company may use such a prediction from the software to determine your insurance premium and benefits	You're easy to get motion sickness, and you notice you seem to get car sick more frequently in autopilot mode	n/a
ML problem type		Regression	Regression	Classification	Classification
Input data type		Tabular data	Tabular/sequential data	Image/video data	Image data
Available dataset		Boston housing [2]	Diabetes dataset [1]	BDD100K [23]	CUB-200 dataset [22]

Table A2: **The four tasks and their explanation goals used in the user study.**

3.2.2 End Users' Explanation Goals

Even for the same user and task, end users' explanation goals, i.e.: the trigger point or motivation to check the explanation of an AI system, may vary in different contexts or usage scenarios. In our study, we aim to capture the fine-grained details of end users' requirements for different explanation goals. We summarize the following potential ***explanation goals*** from prior works [9, 7, 19, 20, 16, 12] as follows:

- **Calibrate trust**: trust is key to establish human-AI decision-making partnership. Since users can easily distrust or overtrust AI, it is important to calibrate trust to reflect the capabilities of AI systems [21, 24].
- **Ensure safety**: users need to ensure safety of the decision consequences [7].
- **Detect bias**: users need to ensure the decision is impartial and unbiased [12, 19].
- **Unexpected prediction**: the AI prediction is unexpected, and/or users disagree with AI's prediction [9].
- **Expected prediction**: AI's prediction aligns with users' expectations [9].
- **Differentiate similar instances**: due to the consequences of wrong decisions, users sometimes need to discern similar instances or outcomes. For example, a doctor differentiates whether the diagnosis is a benign or malignant tumor [12].
- **Learn from AI**: users need to gain knowledge, improve their problem-solving skills, and discover new knowledge [9, 7, 12, 19].
- **Improve the predicted outcome**: users seek causal factors to control and improve the predicted outcome [16, 12, 19].
- **Communicate with stakeholders**: many critical decision-making processes involve multiple stakeholders, and users need to discuss the decision with them [16].
- **Generate reports**: users need to utilize the explanations to perform particular tasks such as report production. For example, a radiologist generates a medical report on a patient's X-ray image [9].
- **Multiple objectives trade-off**: AI may be optimized on an incomplete objective while users seek to fulfill multiple objectives in real-world applications. For example, a doctor needs to ensure a treatment plan is effective as well as having acceptable patient adherence. Ethical and legal requirements may also be regarded as objectives [7].

Each task is accompanied by several explanation goals as shown in Table A2. The tasks and explanation goals were presented in the form of storyboards using graphics and text (Fig. 15).

3.2.3 Creating Prototyping Cards from Explanation Forms

We instantiated the explanation forms as low-fidelity prototyping cards for each task, based on their visual representations. We illustrate this process below:

1. **Create prototyping card templates** We started by creating visual representation templates of the explanation forms. We selected the most common visualizations, based on the summarized visual representations in Section 1. For example, we used bar chart and color map to visualize feature attribution for tabular and image data respectively. Each individual card shows one visual representation of an explanation form. For some explanation forms (such as feature attribution and counterfactual example), we created multiple cards with different variations of their visual representations.
2. **Extract features as content placeholder** We then manually extracted several interpretable features given the AI task. For instance, in the house price prediction task, we extracted features of house size, age, etc. In the self-driving car task, we extracted saliency objects such as traffic signs, road markers, cars, and pedestrians. As a quick prototyping, the feature content may not necessarily reflect the real content generated by XAI algorithms. They mainly serve as content placeholders.

3. **Fill the prototyping templates with content placeholder** The extracted features were then used to fill in the prototyping card templates. The final prototyping cards are shown as figures in Section 1.

After interviewing the first five participants, we revised some prototyping cards based on participants' feedback. For instance, we indicated the position of the input data point on the feature shape and feature interaction cards. We also removed several variations of the cards (such as using a table to represent rules) since participants found them difficult to interpret.

3.3 Study Procedure

The study session consisted of a one-to-one, in-person, open-ended, semi-structured interview and a card selection & sorting. The study procedure has two rounds (Fig. 15). **Round 1** is to familiarize participants with the tasks and explanation goals, and to understand end users' explanation goals for XAI before showing them prototyping cards of the explanation forms. **Round 2** is the selection & sorting to understand users' interpretations and requirements for the explanation forms.

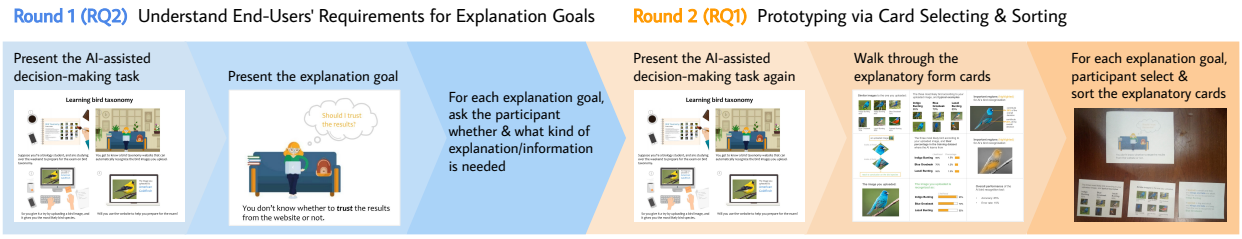


Figure 15: **The user study procedure**. The study consisted of two rounds In the above example, we use the Bird task and the explanation goal of calibrating trust.

3.3.1 Round 1: Interview on Explanation Goals

We began the user study by introducing the researchers and the aim of the study, and went through the study consent form with the participant. The interview started after gaining the participant's written consent.

Task We first introduced an AI-assisted task to participants. The choice of the task was determined by a pre-generated random sequence. The task was presented as a storyboard color-printed on paper. Fig. 16 shows an example of the storyboard of the Health task. The researcher asked the participant to assume s/he was the character in the story context, and went through the task context with the participant by reading the text on the storyboard.

Explanation Goal After confirming that the participant had no questions and fully understood the current task, the research then randomly selected an explanation goal under the task context. The explanation goal was also shown as a storyboard picture, and the researcher read the text on the picture to introduce the explanation goal. Figure 17 gives an example of an explanation goal of unexpected. For each explanation goal, the researcher asked the participants whether they accept AI as decision-support, and need AI to explain its decision. If explanations were needed, the researcher then asked what explanations/further information they request.

After discussing all explanation goals for one task, the participant entered Round 2: card selection & sorting, which is detailed in the next subsection.

After completing one task, if the duration of the interview was less than 30 minutes, the participants were assigned to another task and underwent the same two-round interview procedure. At the end of the interview, the participants filled out a demographic questionnaire (Section 7.1). The study session duration is 67.9 ± 18.8 (Mean $\pm$ SD) minutes (Median: 67 min, Range: 41 - 120 min, each participant's specific data are detailed in Table 6). We audio-recorded the interviews, made observational notes on the card selection & sorting process, and took pictures of the card selection & sorting results. All study materials including the storyboards of tasks and explanation goals, and prototyping cards of explanation forms are listed at the end of this document.

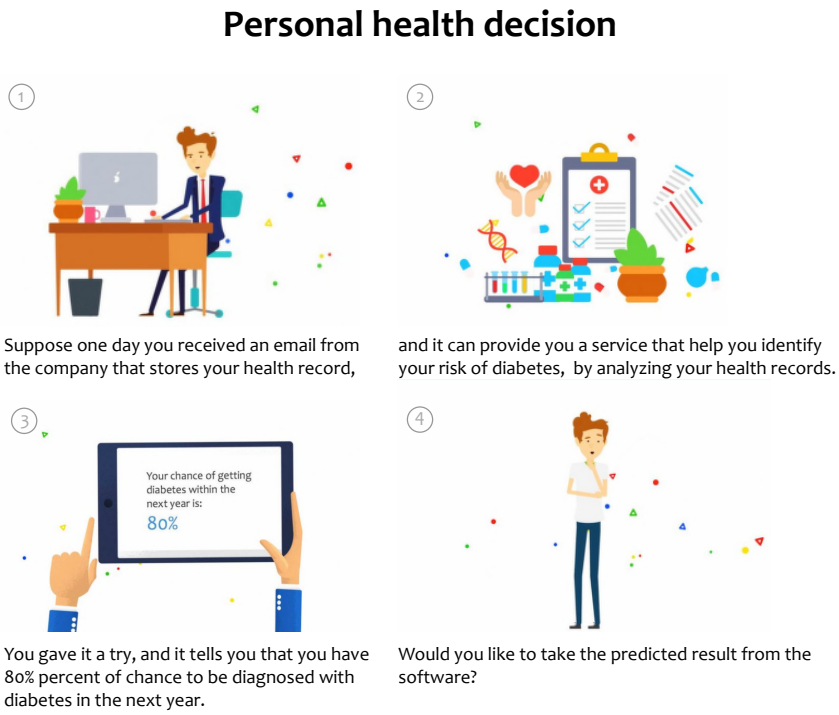


Figure 16: The interview storyboard of the Health task in the user study

Figure 17: The explanation goal of unexpected in the Health task context. The end user may expect to have a high risk of diabetes due to family history. However, AI predicts the risk is only 10% which may not align with the user's expectation.

3.3.2 Round 2: Card Selection and Sorting on Explanation Forms

For each decision-making task, the participants first revisited the task. Then the researcher gave a short tutorial of the explanation forms by walking through the prototyping cards and explaining the information on

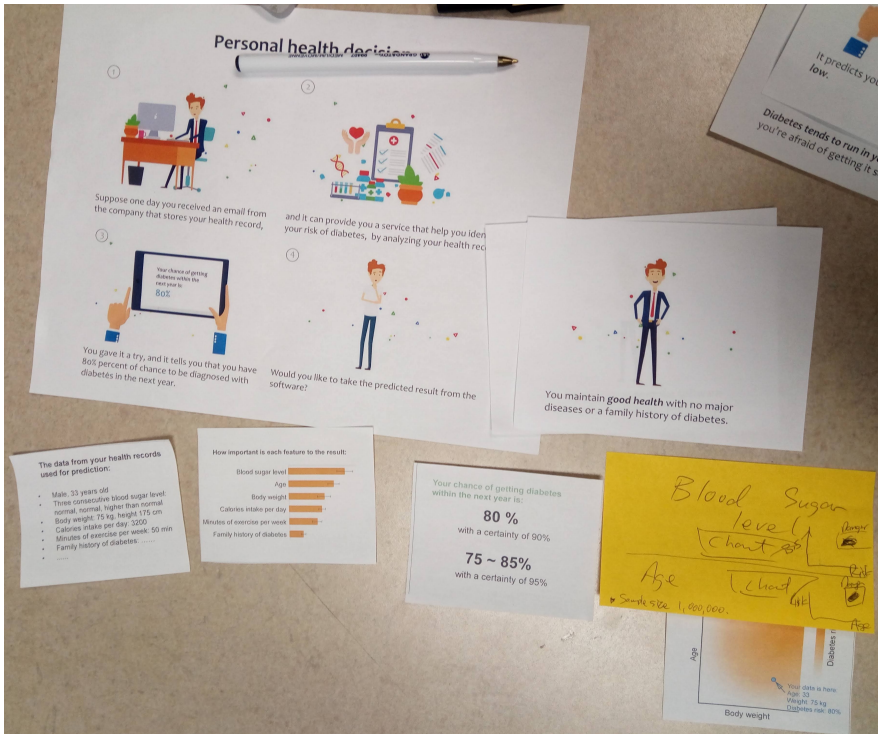


Figure 18: **The card selection & sorting result from one participant**. Given the task and explanation goal, the participant selected and sorted the prototyping cards from left to right according to their usefulness to the given explanation goal. She also sketched to improve the last card on feature interaction.

a card. In this process, the participant could ask questions if s/he did not understand or needed clarification. S/he could also comment on each card. Before moving on to the next step, we asked the participant and made sure they had no questions on these cards.

After confirming the participant fully understand the content of the cards, for each explanation goal, the researcher asked participants to select, rank, and combine the prototyping cards that they found were the most useful ones and could meet their current explanation goals. The participants could comment on any card anytime during this process. They could also modify the existing cards, or sketch on blank cards to create new prototyping cards, and add the newly created/modified cards to the card selection & sorting. After sorting the cards, they were asked to comment on why they selected or did not select a card, and their rationals for making such a sorting. After the card selection & sorting, they were asked whether the combination of cards would fulfill their explanation goals.

3.4 Data Analysis

We utilized a mixed method to analyze the qualitative and quantitative data.

3.4.1 Qualitative Data Analysis

For qualitative analysis, we analyzed the interview data using an inductive thematic analysis approach [4]. A total of 2175 minutes of interviews were recorded and transcribed. We performed coding using Nvivo software. Three members of the research team started with an open coding pass to individually create a list of potential codes. Two additional sets of codes were also applied: 1) the 11 *explanation goals* (listed in Section 3.2.2); 2) the 12 *explanation forms* (listed in Section 1). Upon discussion and applying the affinity diagram process, a unified coding scheme was devised. The coding scheme was not task-specific, as we aim to identify and abstract the general themes despite their different tasks. Two team members independently coded one transcription using this scheme. The first pass of inter-rater reliability Kappa

score was 0.43. After an in-depth discussion with the research team, we further clarified the code definition, merged potential overlapping codes, and removed redundant codes in the coding scheme. The second pass of inter-rater reliability Kappa score was 0.88 on two more transcriptions. The first author analyzed all interview transcripts twice, and the other coder analyzed half of the transcripts.

3.4.2 Quantitative Data Analysis

We performed quantitative data analysis on card selection & sorting of *explanation forms*. To test whether the sorting of explanation forms varies in each condition or has some consistent pattern among explanation goals, tasks, and participants, we conducted the Friedman test on card sorting data. The *null hypothesis* is that there are no cards that are ranked consistently higher or lower than the others. For sorting that showed statistical significance, we further aggregated sortings using Borda count and Instant Runoff Voting[1]. We used an alpha level of .05 for all statistical tests.

We performed clustering analysis to determine the similarity among the 11 explanation goals, and among the 12 explanation forms individually. To cluster the 11 explanation goals, we represented each explanation goal as a 12-dimensional vector, where each number in the vector is the total number of an explanation form card selected for that explanation goal. We then applied k-means clustering on the explanation goal vectors to group 11 explanation goals. We also used principal component analysis (PCA) to reduce the dimension and visualized the relative distances of the 11 explanation goals regarding their card selection similarity. To cluster the 12 explanation forms, we first computed the pairwise similarity matrix measured as the co-occurrence of a pair of cards in card selections. Based on the pairwise similarity matrix, we mapped the 12 explanation forms into a 2-dimensional space using multidimensional scaling (MDS) and visualized it. We then clustered the explanation forms using k-means and hierarchical clustering based on their 2D positions. The statistical and clustering analysis were performed using Python package SciPy and scikit-learn.

3.5 Presentation of Results

The *explanation goals* are marked blue, and *explanation forms* are marked orange. We highlighted key messages using **bold** font. Whenever necessary, we included participants' verbatim quotes despite some minor grammatical errors. Some quotes had their task and explanation goal listed in parentheses.

4 Quantitative Results

A total of 248 valid card selection & sorting data were collected. In the next two sections, we will present quantitative results on card selection and card sorting, respectively.

4.1 Card Selection Results: Preferred Explanation Forms for Each Explanation Goal

For card selection, we analyzed participants' preferences of explanation forms for each explanation goal. The aggregated results are shown in Fig. 20. The ratio of responses is shown as rule (12/15), which means out of 15 card selection responses under a specific explanation goal, 12 selected the explanation form rule. We also visualize the emerging cluster of the explanation goals in Figure 19 based on the vector of the selected explanation forms.

[1] https://pypi.org/project/rankaggregation/

Figure 19: **Clusters of the explanation goals** The explanation goals that are close to each other indicate they have similar patterns on participants' explanation form selection. Specifically, each explanation goal is represented by a 12-dimensional vector, where each number in the vector is the total number of an explanation form selected for that explanation goal. We visualize their relative distances in the 2D scatter plot using PCA dimensional reduction. explanation goals are marked by different colors indicating the cluster they belong to using k-means clustering: **Cluster 1**: Trust, Communication, Unexpected; **Cluster 2**: Safety, Multi-objective alignment, Bias; **Cluster 3**: Expected, Improvement; **Cluster 4**: Differentiation, Learning, Report.

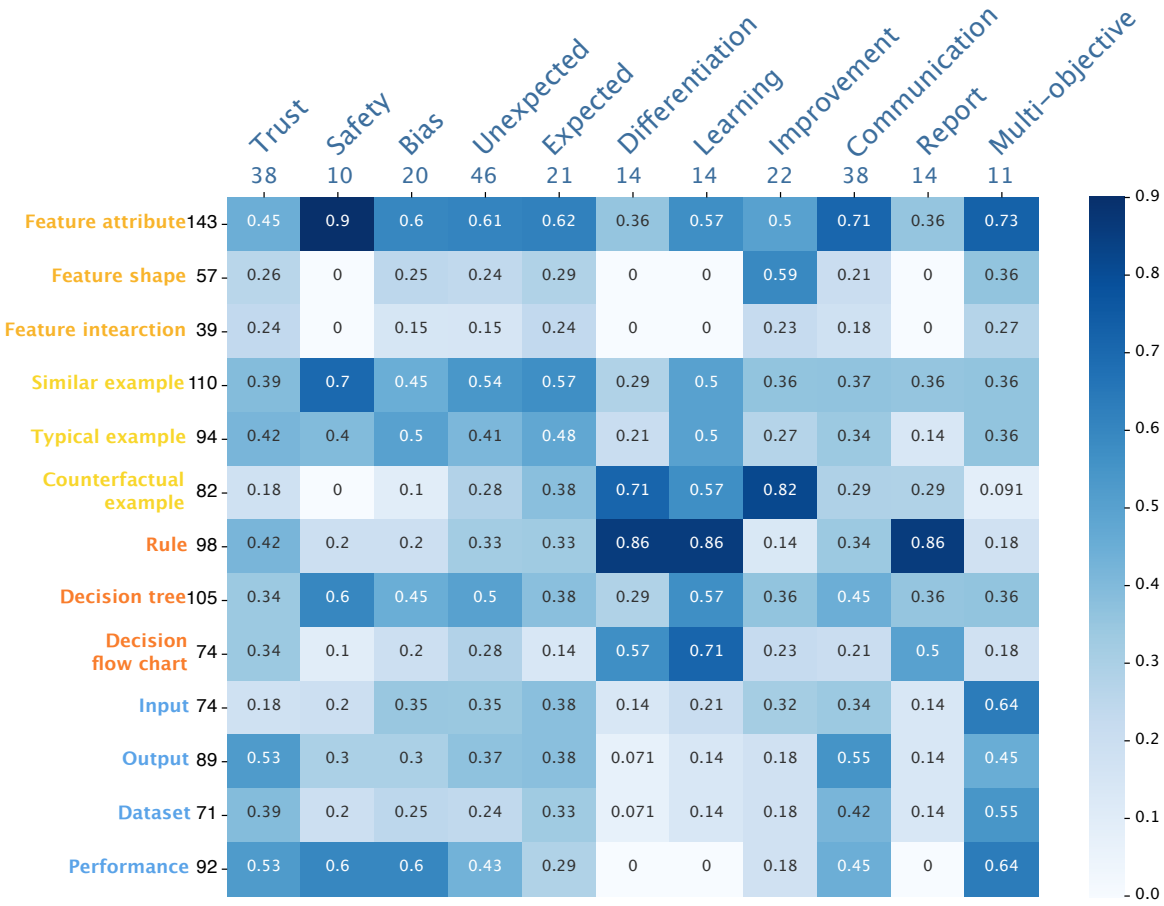


Figure 20: **The explanation form—explanation goal matrix heatmap**. The blueness level and number in the grid is the percentage of an explanation form selected for that explanation goal. The number under each goal (on the horizontal top) is the total number of card selection & sorting data collected for that explanation goal. The number beside each explanation form (on the vertical left) is the total number of times an explanation form was selected in the card selection & sorting data. This is the same figure as in the manuscript, and we display the overall patterns alongside the task-specific patterns.

Calibrating trust

For quantitative results on the most frequently selected explanation forms for the explanation goal to calibrate trust, the top three forms were: performance (20/38), output (20/38), and feature attribution (17/38), which corresponds to the qualitative themes.

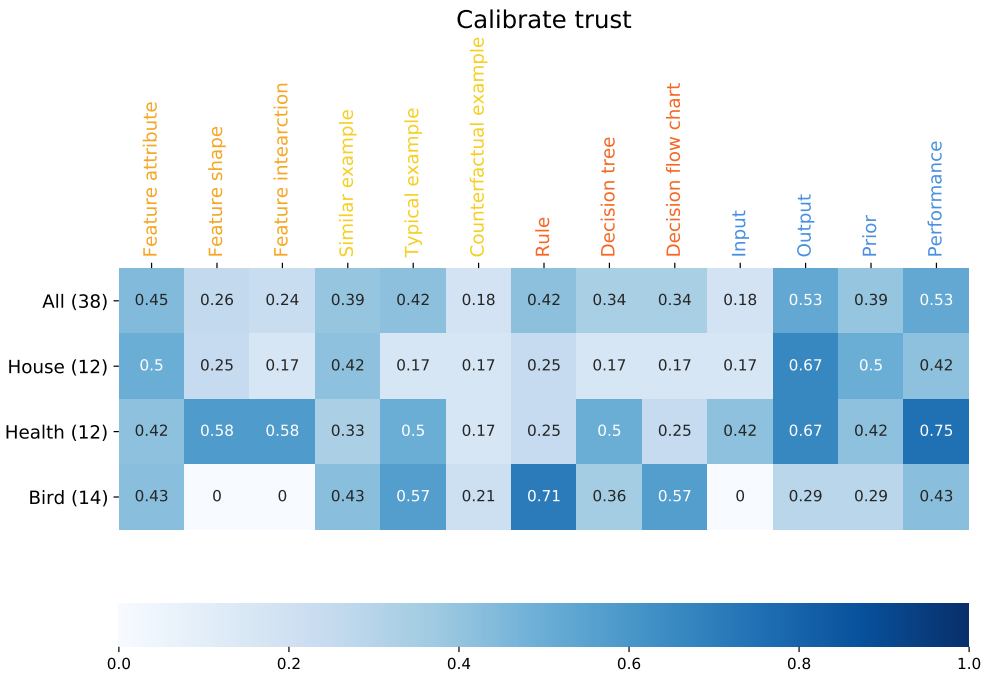
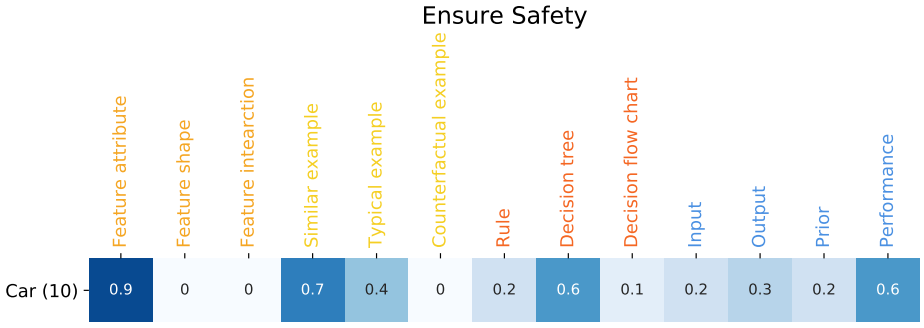


Figure 21: Quantitative result on the explanation goal of calibrating trust. The blueness and number in a cell shows the percentage of an explanation form (column) selected in a task (row). The same for figures below.

Ensuring safety



Regarding the specified information to present AI's testing performance on safety, participants would like to check the objects detected by AI (feature attribution, 9/10):

> *"It shows how it detects the important objects and how it makes decision"* (P03, P05, P27)
>
> *"See if (the feature attributions) align with my own judgment of feature importance."* (P01)

Performance (6/10) were also favourable to check the metrics summary of testing performance. A specified performance analysis in different test scenarios may also help as a safety alert by revealing the weakness of the system.

> *"Let's say I'm driving on a rainy day, then I know that I should be a lot more careful than when I'm with the car in a normal condition."* (P27)

Similar example (7/10) were preferred since it showed *"what's the condition or what kind of decision the car gonna make"* (P32), although participants did not focus on its similarity nature, but rather assumed it

can showcase a variety of cases including the extreme cases. Several participants chose decision tree (6/10) because it *"gave me an overview of how the car makes decision"* (P27).

Detecting bias

A fine-grained performance (12/20) analysis based on protected-feature-defined subgroups [15] can help users to identify potential biases.

> *"I would want to see the certainty and what the prediction error can potentially be for my demographic versus other groups. If it (the prediction error) is quite low, then I would probably worry less about that."* (P22, Health)

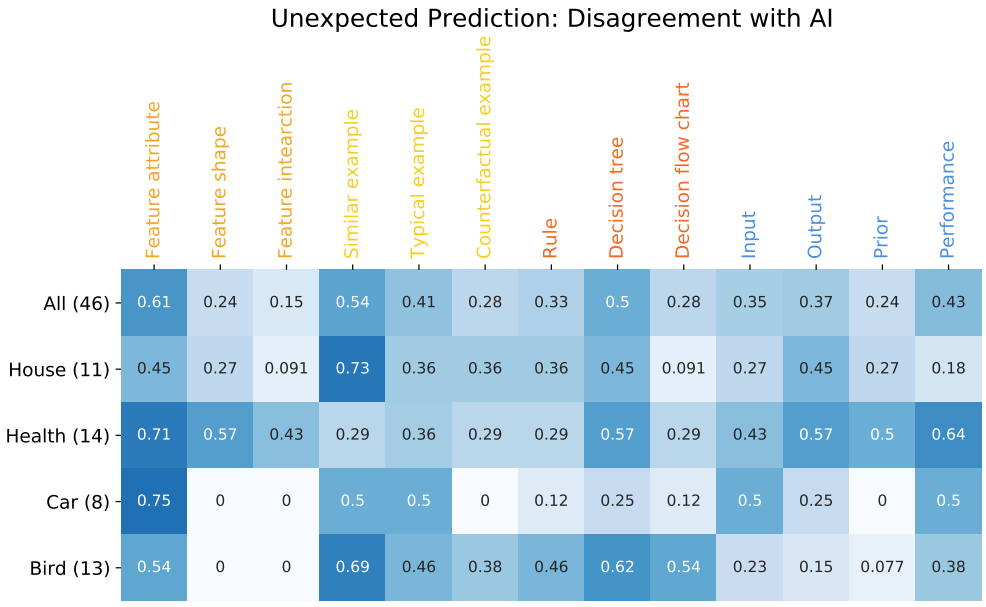
Participants chose similar + typical example (12/20, i.e. out of the 24 card-selection responses on Bias, 12 selected either similar or typical example) to help inspect the data and model, and to compare with other similar instances to confirm their subgroup is included in the model.

> *"You would want to know what the data that it's being drawn from, is it similar to you?"* (P16)

Feature attribution (12/20) was also chosen since participants wanted to check if AI could still detect important features in minority conditions.

> *"I want to see how well AI is performing at night to see what it detected."* (P05, Car).

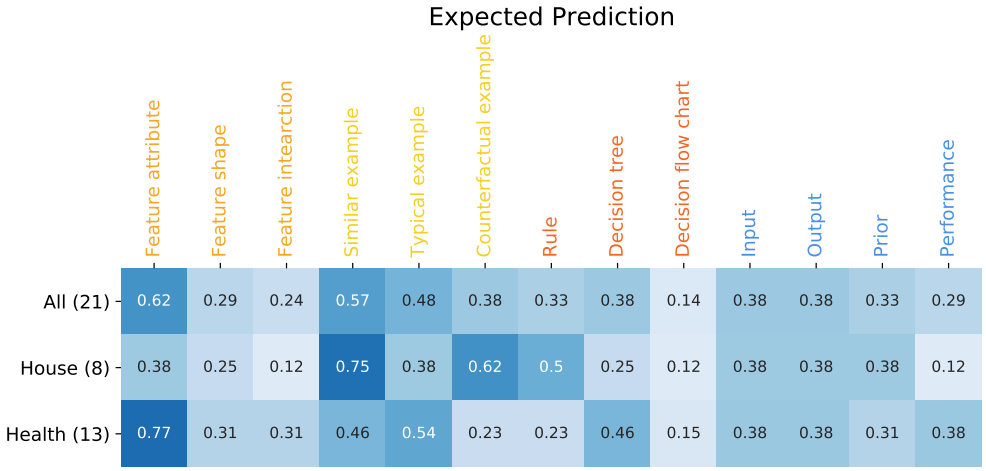
Unexpected Prediction: When Users Disagree with AI



The frequently selected explanation forms are: feature attribution (28/46), similar example (25/46), decision tree (23/46), and performance (20/46).

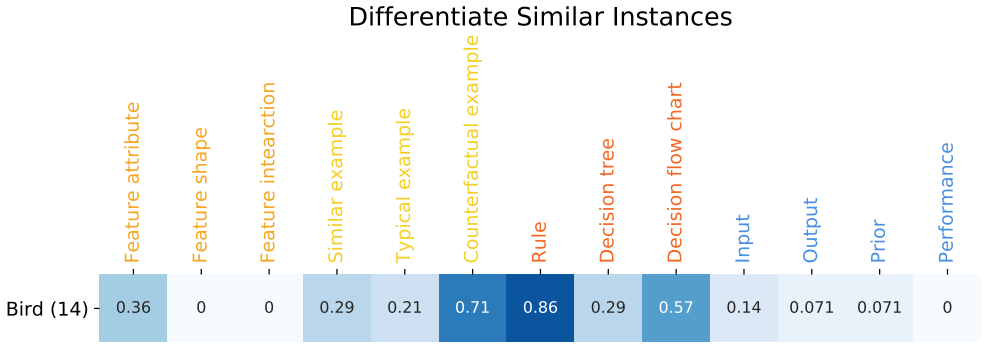
Despite users disagree with AI, if users' judgment is included in AI's differential prediction list or range, users would think AI has the ability to discern similar predictions, and may resolve the prediction disagreement to an extent, as some participants suggested: *"What would be really interesting it's a similar birds list. So if it could provide one or two other possibilities, because then I would know that maybe it thinks it could be a finch, but it's decided it's not a finch (but a Indigo bunting). Whereas if there's no information about other birds, then I would just think of it, 'maybe it doesn't know what it's talking about'"* (P16, Bird task); *"If my prediction appears in similar example, it allows me to judge whether AI is completely unreliable or just need some improvement"* (P01, Bird task). Correspondingly, similar example (25/46) and output (17/46) (listed the top three likely predictions for classification tasks of Bird and Car, or prediction range for regression tasks) were the frequently selected explanation forms for this explanation goal unexpected.

Expected Prediction: When Users Agree with AI



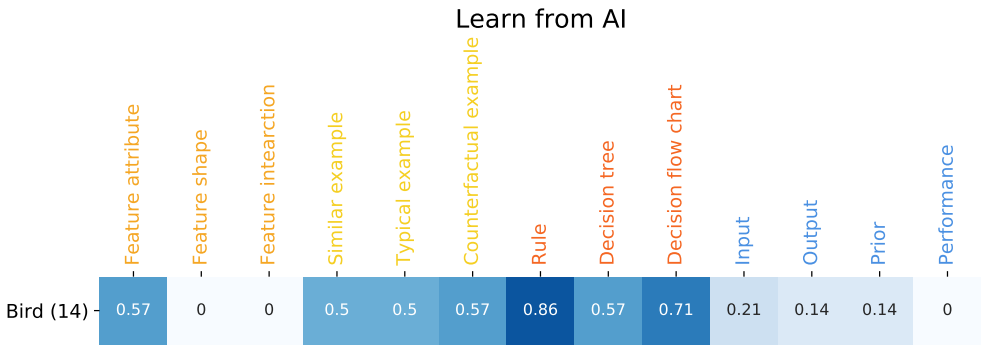
The frequently selected explanation forms are: feature attribution (13/21), similar example (12/21), and typical example (10/21).

Differentiating similar instances



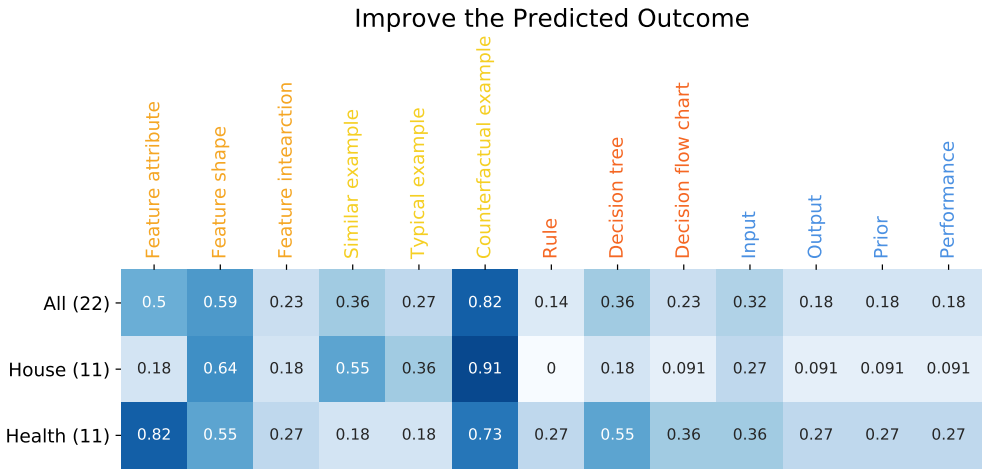
Rule (12/14) and counterfactual example (10/14) were the most preferable forms. Participants chose rule since *"you could write that you differentiated the bird's tail were long or short, or beak thin or thick"* (P10). The counterfactual examples *"identify where specifically to look"* (P16), and *"describe the change, the progress"* (P11).

Learning from AI



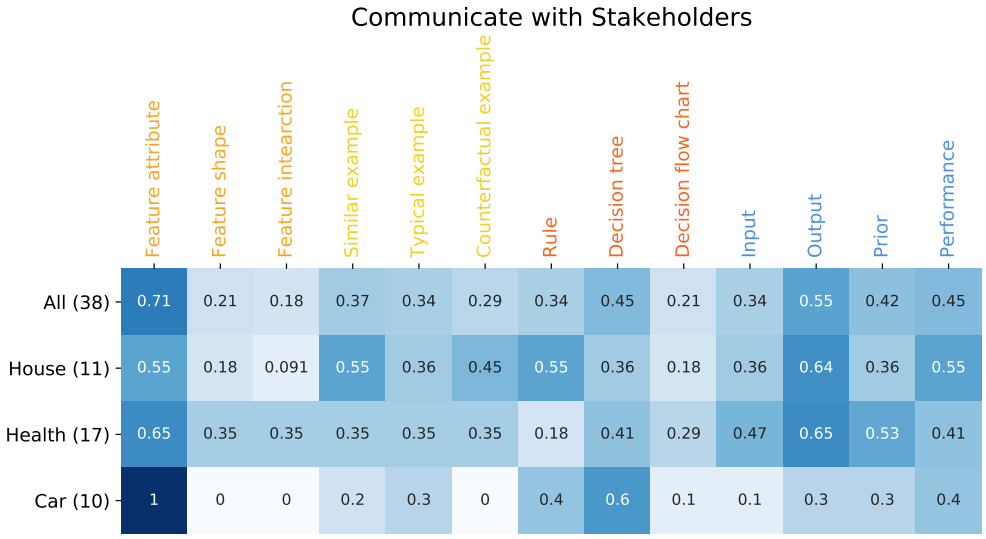
Rule-based explanations (rule: 12/14, decision flow chart: 10/14, decision tree: 8/14) were more favourable for the explanation goal to learn, since they showed *"a learning process. It has like how you could recognize a bird. So help me to learn some new knowledge"* (P02). *"(decision tree) includes the big tree of the birds. I can just choose which bird I want to know, and I will know their relationship and their differences"* (P11). Same as in Report, participants would prefer to see *"the graphics and text combined"* (P02): *"It combines text and pictures, and they are relevant to each other. It's kind of a multi-modal learning"* (P04).

Improving the predicted outcome



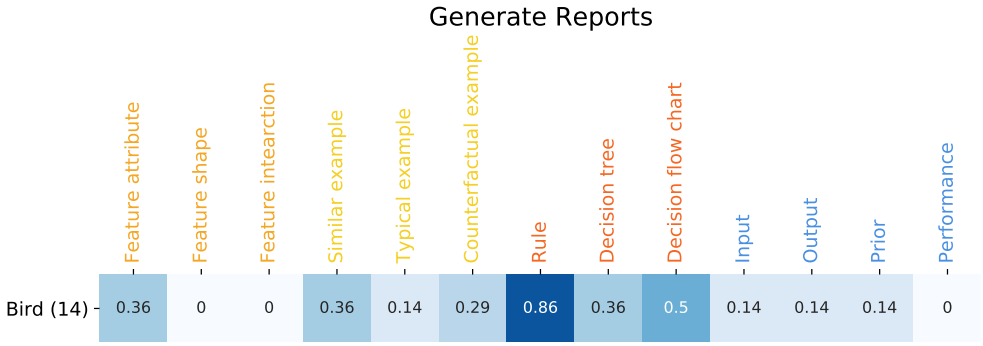
Counterfactual example (18/22) and feature shape (13/22) were the top two selected forms. While counterfactual example provides how to achieve the target outcome change by adjusting the input features (counterfactual reasoning), feature shape (and feature interaction) allow users to adjust features and see how that leads to outcome change (transfactual reasoning [10]).

Communicating with stakeholders



While output (21/38) and performance (17/38) provide AI's result and help to build trust, feature attribution (27/38) and decision tree (17/38) show the breakdown factors and internal logic behind the prediction.

Generating reports



Rule(12/14), decision flow chart(7/14), and feature attribution(5/14) are the most frequently selected explanation forms.

Rule were selected because its text description format can conveniently generate text reports.

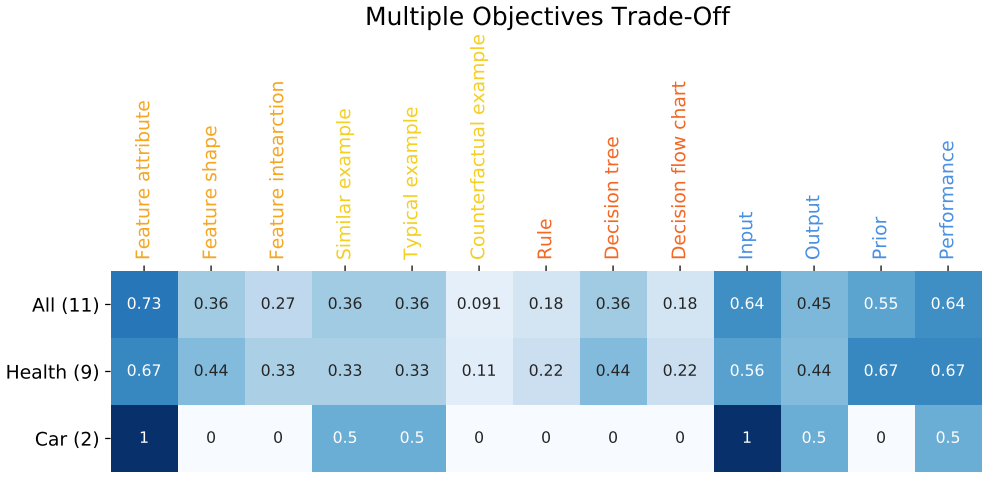
> *"I have to write the explanation"* (P08, P09); *"You can not only by looking at the images and get some explanation. You need some more specific description."* (P08)

In addition, adding image to the text *"would be complementary"* (P10) to each other, and the format of image + text were more favourable by many participants.

Feature attribution and decision flow chart are the second most favourable explanation forms since they both highlight features and were presented as image format (in the bird recognition task).

> *"Rule is just describing and writing. It doesn't really show you a visual on how to compare them."* (P06)
>
> *"Feature attribution and decision flow chart (presented in image format on bird recognition task) highlights what rule is saying, this knowledge complements your statement."* (P10)

Multiple objectives trade-off



Participants' choices of the explanation forms were widely distributed, and they wanted as much information as possible, *"I want all the data"* (P23). In particular, participants chose explanations related to AI model's performance metrics, such as performance (7/11), input (7/11), dataset (6/11), and output (5/11). Feature attribution (8/11) were also preferable to *"get re-evaluated based on the important features"* (P16), *"I would want to know what factors in feature attribution, how have leads way in each of them"* (P22).

To gain an intuitive understanding of how similar the explanation forms are to each other, we applied multidimensional scaling and visualized their similarities as distances on a 2D scatter plot shown in Figure 22, as well as a dendrogram and pairwise similarity heatmap in Figure 23. The similarity was measured as the co-occurrence of a pair of explanation forms in card selection. Based on the 2D positions of the explanation forms, we applied k-means and hierarchical clustering analysis and yielded similar clustering patterns.

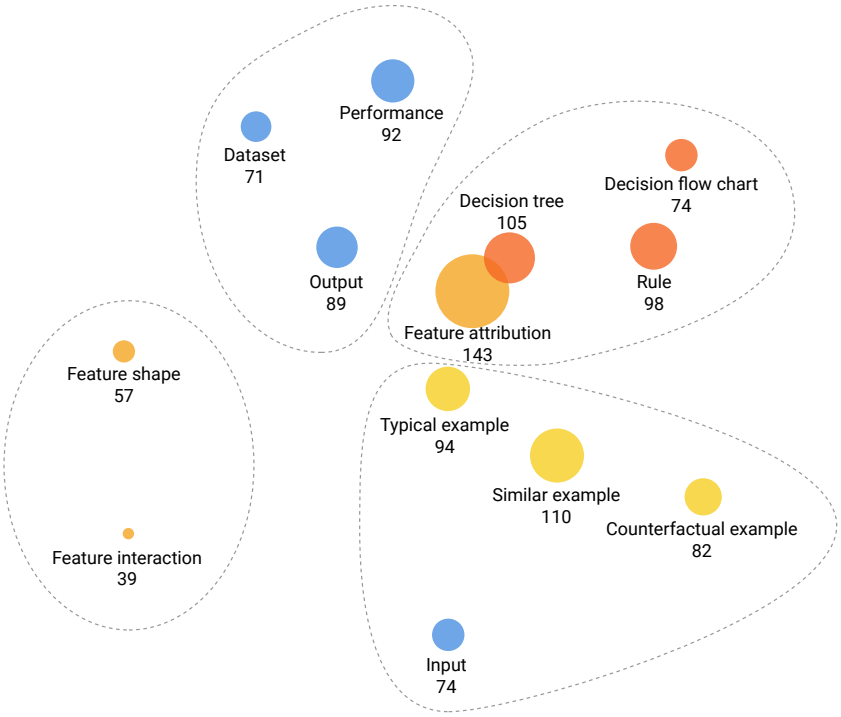


Figure 22: **Visualizing the similarities of the 12 explanation forms**. The explanation forms that are close to each other indicate they are more likely to be selected together to construct an explanation. The total number of times an explanation form was selected is indicated as the number below its name, also proportional to its dot size. The dot color indicates its category: feature, example, rule, and contextual information. The categorical clustering roughly corresponds to the automatic k-means clustering (circled together with a dotted line), which is based on the pairwise similarity matrix measured as the co-occurrence of a pair of cards in card selections from the user study.

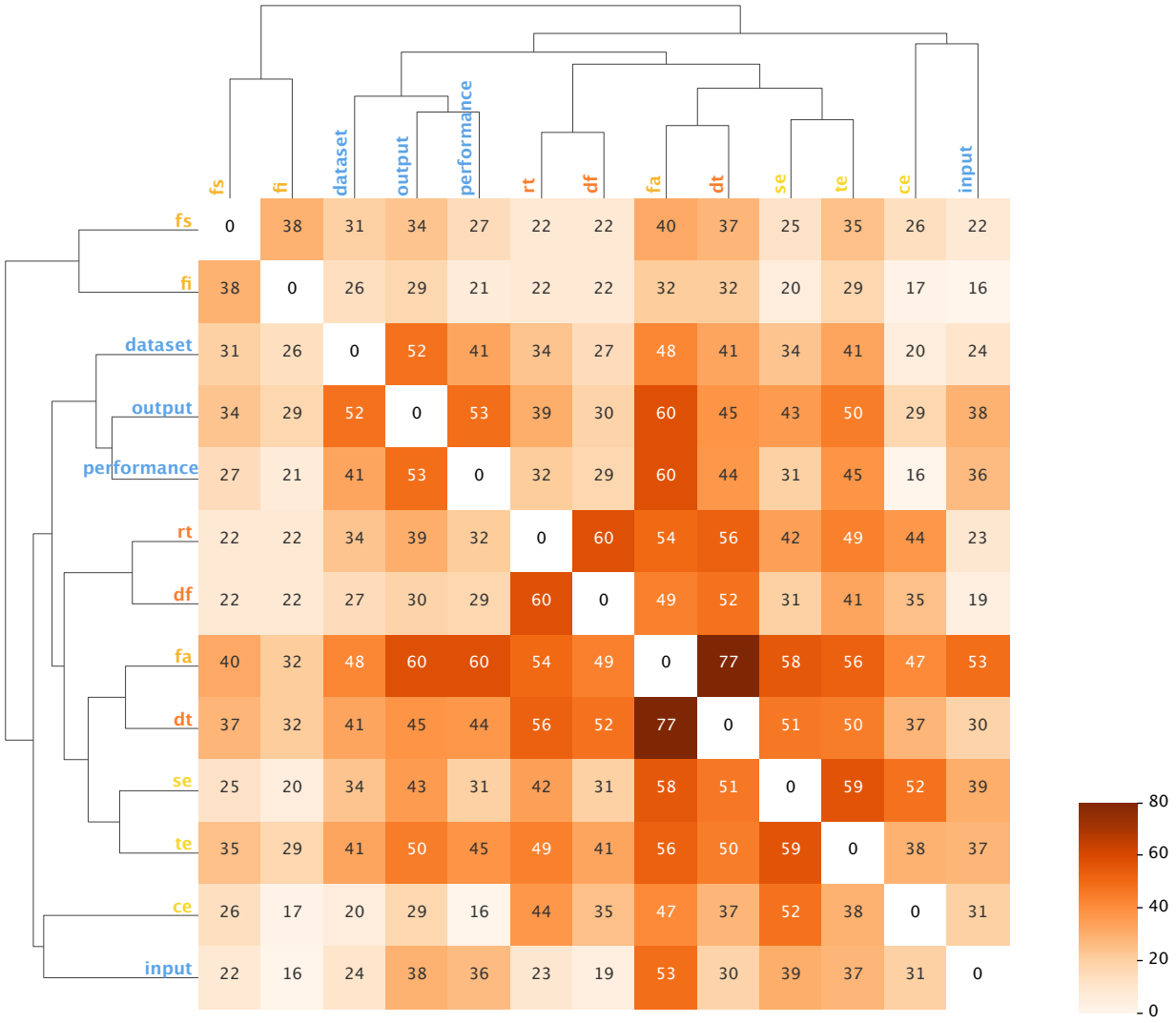


Figure 23: **Similarity matrix and dendrogram of the 12 end user-friendly explanation forms**. The pairwise similarities are measured by the co-occurrence of two cards selected in a card selection & sorting response. Darker orange indicates the two pairs are more likely to be selected in the same explanation form card combination, and the co-occurrence numbers are shown in each grid. The dendrogram was generated using hierarchical clustering.

4.2 Card Sorting Results on Explanation Forms

We performed Friedman tests to see if there is any significant difference (i.e., consistent pattern) of the card sorting distribution among the 11 explanation goals, 4 tasks and 32 participants.

For the 11 explanation goals, except for the explanation goal of expected that has no consistent pattern for the sorting of explanation forms, the rest of the explanation goals had a consistent pattern ($p < 0.05$). Table A3 summarizes the aggregated sorting for explanation goals that have significant consistent patterns.

Table A3: **Aggregated sorting of explanation forms for various explanation goals**. The number following each explanation goal shows the total number of collected sorting data for that explanation goal. The number after each explanation form indicates its number of times being selected for a particular explanation goal.
fa: feature attribution; fs: feature shape; fi: feature interaction
se: similar example; te: typical example; ce: counterfactual example
rl: rule; dt: decision tree; df: decision flow chart

Explanation Goal	**Mean Ranks of explanation forms**
Trust 38	performance 20, fa 17, output 20, te 16, rl 16, se 15, dataset 15, df 13, dt 13, input 7, fs 10, fi 9, ce 7
Safety 10	fa 9, dt 6, se 7, performance 6, te 4, prior 2, input 2, output 3, rl 2, df 1
Bias 20	fa 12, performance 12, te 10, se 9, dt 9, input 7, dataset 5, df 4, fs 5, output 6, rl 4, fi 3, ce 2
Unexpected 46	fa 28, se 25, dt 23, performance 20, te 19, input 16, output 17, df 13, rl 15, ce 13, dataset 11, fs 11, fi 7
Expected 21	No sorting patterns are statistically significant
Differentiation 14	rl 12, ce 10, df 8, fa 5, se 4, dt 4, te 4, dataset 1, input 2, output 1
Learning 14	rl 12, df 10, fa 8, ce 8, dt 8, se 7, te 7, input 3, dataset 2, output 2
Improvement 22	ce 18, fa 11, fs 13, dt 8, input 7, se 8, df 5, te 6, performance 4, fi 5, dataset 4, rl 3, output 4
Communication 38	fa 27, output 21, performance 17, dataset 16, dt 17, se 14, input 13, te 13, rl 13, ce 11, df 8, fs 8, fi 7
Report 14	rl 12, se 5, df 7, dt 5, fa 5, ce 4, te 2, input 2, dataset 2, output 2
Multi-objective alignment 11	fa 8, performance 7, input 7, dataset 6, se 4, te 4, output 5, dt 4, fs 4, df 2, rl 2, fi 3, ce 1

For the 4 tasks, the explanation form card sorting on all 4 tasks showed some consistent patterns regardless of their varying explanation goals. The aggregated sorting for the four tasks are shown in Table A4.

Table A4: **Aggregated sorting of explanation forms for four tasks in the user study**. The number after each task shows the total number of collected sorting data for that task. The number after each explanation form indicates its number of times being selected for a particular task.
fa: feature attribution; fs: feature shape; fi: feature interaction
se: similar example; te: typical example; ce: counterfactual example
rl: rule; dt: decision tree; df: decision flow chart

Tasks	Mean Ranks of explanation forms
House 53	se 31, output 24, fa 22, ce 26, input 15, te 17, dataset 17, dt 15, performance 15, fs 17, rl 17, df 7, fi 7
Health 86	fa 56, input 38, performance 44, dataset 39, fs 40, te 36, dt 40, output 44, se 30, df 22, fi 32, rl 20, ce 26
Car 40	fa 47, performance 22, dt 20, se 18, te 15, input 11, output 10, rl 9, dataset 5, df 5,
Bird 69	rl 52, df 40, se 31, ce 47, fa 46, dt 30, te 26, performance 11, output 11, input 10, dataset 10

For the 32 participants, over half 59% (19/32) of the participants demonstrated some consistent patterns of sorting the explanation forms, despite different tasks and explanation goals ($p < 0.05$). The chi-square test showed there is no statistically significant association between gender and the explanation form card selection; whereas card selection preferences do differ by age group, educational level, familiarity with AI, and attitude towards AI ($p < 0.05$).

5 Detailed Qualitative Results

For each explanation form, the user study (in Round 2) identifies users' interpretations including pros, cons, applicable explanation goals, and users' requirements shown as design implications.

5.1 Feature Attribution

5.1.1 Pros

In the study, we used a bar chart to represent feature attribution or importance score for tabular data, and color map and bounding box object detection for image data (Fig. 1). All participants **intuitively understood feature attribution**, and over half selected it (143/248) and ranked it at a relatively top position.

> *"Feature attribution uses a simple way to highlight the most important parts, and you can see very clearly at your first sight how this can be recognized."* (P04, Bird, Learning)
>
> *"It's easy to read. ...And you have a bar (chart) here it's really clear information that people understand instantly."* (P28, House, Trust)

By showing *"finer details"* (P10) and *"breakdown and weights of features"*(P23) *"that AI took into account"* (P31), participants perceived feature attribution **can answer *"how"* and *"why"* questions**.

> *"tells me why"*(P20), *"gives me the behind the scenes"* (P24), *"tells me how AI read things and how it makes decisions"* (P03), *"have an understanding of how much weight AI is giving to each of the factors"* (P22), and *"identify key aspect, ...support its reasoning"* (P18).

5.1.2 Applicable Explanation Goals

By checking the ranking of feature importance score, participants would instantly *"compare with my own judgment, to see if that aligns with my feature ranking"* (P01, Car, Safety), especially when participants need to **verify AI's decision**.

5.1.3 Cons

Although a causal relationship may not be confirmed, some participants tended to **assume feature attribution is causal**, or simplify the relationship among features by assuming they are **independent of each other**. This was usually occurred when they were seeking explanation to improve the predicted outcome. Furthermore, participants were likely to be informed by the feature importance score to prioritize the most important features to take actions upon.

> *"Seeing that body weight is more important than exercise, I think I will focus on changing what I ate, instead of like responding by going to the gym everyday."* (P16, Health, Improvement) – Relies on feature attribution to improve the outcome.
>
> *"It (feature attribution) shows what are the most important factors that AI has taken into account, so you could target the biggest factors."* (P31, Health, Improvement) – Assumes a causal relationship and prioritizes her action accordingly.
>
> *"If my blood sugar puts me at a super high risk here, but my caloric intake doesn't actually put me at that higher risk, it's like a lower risk, then I would rather just focus on blood sugar. "* (P22, Health task) – Ignores the complex interaction between blood sugar level and caloric intake.

5.1.4 Design Implications

To avoid the above causal illusion [14], The XAI system design may need to **alarm users** (explicitly or implicitly) that changing the important features may not necessarily lead to the outcome change in the real world, because most AI models can only capture the correlation between features and prediction, and correlation does not necessarily imply causality.

For the prototyping card design, the AI researcher may consider **varying different visual representations** of the feature importance, such as showing the feature ranking only and allowing users to check the detailed attribution scores on demand, or allowing users to set a threshold on the attribution score and only showing features above the cut-off value, as suggested by a few participants.

> *"If the percentage (of the feature) is below the cutoff value, the users does not need to see (the feature), reduce the cognitive load."* (P04, Bird, Learning)

5.2 Feature Shape

5.2.1 Pros

Participants **liked its graphical representation** of showing the relationship between one feature and prediction.

> *"It (feature shape on exercise and diabetes risk) feels so easy to latch onto like it's something that you can impact and something that's very tangible."* (P22, Health, Trust)

5.2.2 Applicable Explanation Goals

The slope of the curve in feature shape line chart allows users to easily check how changing one feature would lead to the change of the outcome. Thus, many participants intuitively used feature shape for **counterfactual reasoning**, especially to improve the predicted outcome.

> *"I would be interested to see how much like here (feature shape) increasing the exercise by a small amount actually makes a really big difference. So that's also helpful to decide what you should be*

focusing on to try to avoid it (diabetes). The shape of the curve actually helps. Coz if I was out here [pointing to the flat part of the curve], then it would not be as helpful for me to increase my exercise." (P16, Health, Improvement)

By showing the relationship between the protected feature (such as gender, ethnicity, [15]) and outcomes (such as loan approval), the feature shape explanation is also helpful to reveal potential bias, i.e.: to check if different assignments of the protected features (male, female) will lead to prediction differences.

"If these features are related to diabetes, then it (AI) should present some (feature shape) cards to tell me if the gender, age and ethnicity (will affect diabetes prediction), so this image (feature shape) would be really helpful." (P02, Health, Bias)

5.2.3 Cons

One drawback of feature shape pointed out by a few participants is that it does **not considerate feature interactions**.

"This one (feature shape on house size and price) is not based on the bigger the house, the higher you can sell, because it is based on a lot of features. Let's say the house is 2000 square feet. It was built in 1980. Another one is 1000 square feet, but it's just built a decade ago. So its (the latter) price will be much higher than this one (the former). You cannot just base on a house area and then determine the price." (P30)

Another drawback is, since a feature shape plot can only explain one feature at a time, to present explanation for multiple features, the interface will be occupied by multiple feature shape plots which may –

*"make your page so **overloaded**, so people just get tired. You want to make it as clear as possible. So if (there is) some unnecessary information people just intimidated."* (P28)

5.2.4 Design Implications

One suggestion for the above weaknesses is that feature shape can be accompanied by other explanation forms and **show on-demand**. Users can select their interested features from a feature list, or click a feature from other explanation forms (such as feature attribution, counterfactual example or rule), and view feature shape plots of selected features, as participants suggested:

"If I can click on this (feature attribution) and then I can get this chart (feature shape), I think that would be good. I don't think everyone is going to click it, but I think (if) people want more information, you will click it." (P20, House)

Many participants tended to check the **local position of their input data point** on the global feature shape diagram.

"It's good to see where exactly on a (house price) scale you are." (P20, House, Trust)

And P30 suggested feature shape could have the assumption that for all the other features that are kept constant, they should be as similar to user's input features as possible.

"The AI should assume all the other features are almost the same as mine, considering this hypothesis then this is the (feature shape) curve".

5.3 Feature Interaction

5.3.1 Applicable Explanation Goals

Since feature interaction just adds one more feature to the feature-outcome plot to show feature-feature interactions, it can be regarded as an expanded version of feature shape, and many of the above findings on feature shape apply to feature interaction as well. Similar to feature shape, feature interaction also supports **counterfactual reasoning** by including two or more features instead of one (as in feature shape).

"(feature interaction on age-body weight interaction) If you put yourself in a hypothetical guessing, you're in this age and this is your body weight, and you can already tell the chances (of diabetes) are high." (P23, Health, Trust)

5.3.2 Cons

Since it adds feature interactions, four participants found *"the graph is **less accessible to understand**"* (P22).

5.3.3 Design Implications

Similar to feature shape, participants would like to **choose their interested feature pairs** to check their interactions on the feature interaction diagram. Since the combination of features is large, the XAI system may be able to **suggest interesting feature interactions and prioritize** the feature pairs which have significant interactions.

> *"If I click on any two of them (features), show the relationship between them. If I can choose age and blood sugar level, then probably there is some correlation between them. If it is statistically significant, then I would want to know that. If there is no significance between, for instance, age and body weight, then I don't think it should tell me that. If the AI can tell me that this combination really is important for you to look into, then the priority would also make a lot of sense."* (P23, Health, Unexpected)

5.4 Similar Example

In our study, most participants regarded both similar example and typical example as similar examples. Only a few participants got the idea that with a typical example, *"you're getting the average"* (P20). Thus in this section, we state the themes on similar example as well as the common themes of similar and typical example.

5.4.1 Pros

Participants **intuitively understood** the concept of similar example. Similar example uses analogical reasoning to facilitate user's sense-making process.

> *"It just intuitively makes sense to me. ...similar and typical example are much easier. I don't have to think about them before figuring it out."* (P16, Bird, Trust)

> *"(similar and typical example) It's similar to how humans make decisions, like we compare similar images to the original (input) one."* (P02, Bird, Trust)

5.4.2 Applicable explanation goals

Unlike other explanation forms that reveal AI's decision-making process (such as rule-based explanations) , *"even though these (similar and typical example) aren't much specific about how it's actually doing the (decision) process"* (P16), participants' minds automatically made up such a process by themselves by **comparing instances**. Such comparison mainly allow users to **verify AI's decisions** and to calibrate their trust. The common explanation goals for which similar example were selected are:

1) To build trust, especially from a personal and emotional level.

> *This (similar example) made me trust on an emotional level. Because I'm thinking, 'Oh really? I am only 33 years old.' Like I probably not going to get diabetes. But then I'm reading about somebody that does (get diabetes), that sounds a lot like me, it kind of emotionally makes me feel like, 'Oh geez, maybe it is accurate.' So this (performance, output) is like using my brain, and this one (similar example) kind of got me in the gut like, 'Oh, okay. This could actually happen to me. It happened to this person who sounds a lot like me.' "* (P16, Health, Trust).

2) To verify the decision quality of AI.

> *"It's like a proof for my final decision."* (P30)

> *"Because AI has only 85% accuracy, I want to see similar ones, and what AI thinks they are."* (P14, Bird, Trust)

"If it doesn't align (with my prediction), then I want to see some similar houses to remake the judgment."(P04, House, Unexpected).

3) To assess the level of disagreement when AI made an unexpected prediction, and to reveal potential flaws of AI.

"If my prediction appears in (a list of) similar examples, it allows me to judge whether AI is completely unreliable or just need some improvement." (P01, Bird, Unexpected)

5.4.3 Cons

Showing examples for comparison may not be applicable when input data is incomprehensible or **difficult to read and compare**.

"I think (similar and typical example) it's not important to me. Because I need to read other people's status, read their records." (P02, Health, Trust)

In addition, participants easily got confused when instances in similar example have **divergent predictions**. This problem might be solved by typical example which is stated in Section 5.5.

"(similar example) It's not really telling you if it (the input) is the one (prediction), so it could be this (prediction) or this or this [pointing to different predictions on similar example card]." (P26, Bird, Trust)

"This one (similar example) has too many choices (predictions), it's too confusing."(P05, Bird, Trust)

5.4.4 Design Implications

As mentioned above, participants had to compare the features in similar example by themselves. It is important for the XAI system to support such **side-by-side feature-based comparison** among instances such as input, similar, typical, or counterfactual example, especially when the input data format is difficult to read through.

"I don't want to read the text (in similar and typical example), it is better to show those features and examples in a table for me to compare directly, also highlight the important features as an analysis process." (P29, Health, Trust)

" Maybe it could help the doctor to pinpoint things that are similar or different between these cases."(P31, Health, Communication)

"I would like a comparison. That's my own house (input), which probably will be off the top somewhere. And I'm comparing it with other information (typical example and counterfactual example). So in a column, and I can compare it. For the layout, maybe you can do a product comparison." (P03, House, Expected)

5.5 Typical Example

5.5.1 Pros

One drawback of similar example is that it may make users confused about similar data instances, especially when they have different predictions. Typical example may overcome this problem, since the typical examples for different predictions are more **distinct and separable** than the nearest neighbors of similar examples.

"(typical example) You actually made a category of each one. I remember in cognitive psychology, there's a theory. I don't remember the name, but if you clearly separate each category, that helps people to differentiate the different categories, then remember. But for this one (similar example), you have to read every one (instance) of them." (P04, Bird, Learning)

5.5.2 Applicable explanation goals

Since typical example represents the typical case for the outcome, it may help to reveal class-specific characteristics or even potential problems in the AI model or data, for example to **reveal bias**.

> *"If I'm concerned about what group the data is coming from, I would love if the typical case like the average that comes up says like, male, this age, and the factors were quite different from mine, then I kinda go, 'huh?' But if it could give me a typical case that's actually quite similar to me, then I would be less worried about it not performing well with my group."* (P22, Health, Bias)

Unfortunately, most participants did not realize the meaning of typical example and did not make use of such "debugging" property.

5.5.3 Design Implications

In addition to showing typical examples from different predictions (between-class variation), in some cases, it might be beneficial to show **different variations of typical example** for a particular prediction (within-class variation).

> *"It's showing different pictures of the same bird, and the colors even look different. So it's saying maybe, 'Oh, I get it, we have the male and female.' So it's showing different looks that the bird can have."* (P06, Bird, Learning)

Contrary to typical example, some participants expected to see non-typical or edge cases that represent rare but severe consequences, mainly due to safety and bias concerns.

> *"So they (similar and typical example) don't really provide enough information about when the weather is different and when you're driving at night, the results from non-typical conditions."* (P27, Car, Bias)
>
> *"I still don't know if the dog jumps out of nowhere. so maybe the (similar example) similar traffic conditions can see the extreme cases."* (P03, Car, Safety)

5.6 Counterfactual Example

5.6.1 Pros and Applicable Explanation Goals

In our study, counterfactual example was shown as two instances of different predictions, with their feature differences highlighted while keeping other features the same (Figure 7). This visual representation of counterfactual examples can serve for different explanation goals depending on the task context. In predictive tasks (House and Health), participants regarded counterfactual example as the most direct explanation form to **suggest ways to improve the predicted outcome**.

> *"For renovations, I think that's (counterfactual example) the only card I would choose. The only one that really tells me that I can do something to increase the price."* (P20, House, Improvement)

Whereas in recognition task (Bird), counterfactual example is suitable to **show the differences** to differentiate two similar predictions.

> *"Counterfactual example let me learn their relationship, highlight the difference between the two (birds). Help me remember the different features."* (P11, Bird, Learning)

5.6.2 Cons

Some participants did not understand the meaning of counterfactual example, and could not capture the nuance between feature attribution and counterfactual example, since they both have features highlighted but for different reasons: feature attribution highlights features that are important for prediction, whereas counterfactual example highlights what features need to change for the potential outcome to happen.

Counterfactual example may have the drawback of making users **confused about similar instances**, especially in recognition tasks.

> *"I think this tool (counterfactual example) will make me remember the wrong thing. I'm already confused. It shows information that is similar."* (P11, Bird)

Thus, it may not be the beginning explanations and may only show up on-demand, for example, for the two explanation goals of improvement and differentiation mentioned above.

5.6.3 Design Implications

The two **contrastive predictions in a counterfactual example can be user-defined or pre-generated** depending on the specific explanation goals. One prediction is usually from user's current instance such as input, and the alternative prediction could be: *"the next possible prediction"* (P18, Bird, Report), users' own prediction when there is a disagreement (unexpected), the prospective prediction to improve the predicted outcome, and the easily confused prediction to differentiate similar instances.

The generating of counterfactual features may also receive user-defined or pre-defined constraints, such as: **1**) constraints on the counterfactual feature type to include **controllable features only**' **2**) generate **personalized counterfactual suggestions** based on features that users look upon: *"the recommendation should be a lot based on what I do"* (P24, Health); and **3**) constraints on the range of specific counterfactual features: *"AI should accept my personalized constraints on budget* (P01, House, Improvement). Given these constraints, the XAI system can also provide multiple improvement suggestions for users to choose from (P01, P11), and may give weights or relative rankings on multiple suggestions.

5.7 Decision Rule

Many participants noticed different forms of rule-based explanations (rule, decision tree) provided *"basically the same information"* (P02, Health), *"all show the decision process"* (P10, Bird), and were only different in the text (rule) or graphical (decision tree) representation.

5.7.1 Pros

Several participants regarded rule can ***"explain the logic** behind how the AI makes decisions"* (P27). Particularly, the text description format is *"like human explanation"* (P01, House, Trust), and *"simple enough and understandable"* (P11).

5.7.2 Applicable Explanation Goals

The above pros make rule suitable for verbal and written **communications**. Text format may also help to dispel confusion, since some participant regarded texts as being more precise than images, thus **facilitate learning**.

> *"In this case (Bird, Unexpected), I don't want to see the highlights (feature attribution). I want it to see points, the specific parts and give me some explanation. If I'm trying to prove myself wrong, or if I want to see how AI system can prove me wrong, I want to see more precise text, and precisely point out the important information."* (P04)

> *"The written helps because it's more exact, whereas the pictures, ...the blue in the picture might not be the blue that was in the written."* (P05, Bird, Learning)

> *"(rule) It's listing out something that a person might miss in the picture."* (P18, Bird)

However, when the input is image data, some participants also mentioned providing text explanations only was not enough.

> *"(Rule) It doesn't really show you the bird that you were looking at. Lots of birds have small thin bills short tails...if I can't see a picture of it, then it's not as helpful."* (P06, Bird, Trust)

And many participants suggested *"ideally you'd want both written and pictures"* (P05) to complement each other.

5.7.3 Cons

Rule is **very sensitive to the degree of complexity** in text descriptions, as an increase in rule length or number of features will dramatically reduce its simplicity and the above advantages [18]. However, if the rule clauses are short, the explanation may not be precise and satisfying as well, as P06 pointed out,

> *"It (rule) is just too broad, it could apply to so many other birds."*

Another concern is that since participants lack technical knowledge, some of them misinterpreted rule as instructions human fed to the AI.

> *"(Rule) it is giving very clear instructions to the AI, like written text instructions, these are already fed into the system."* (P09, Car, Safety)

5.7.4 Design Implications

To reduce the cognitive load of complex rules, a few participants suggested **trimming the rules** by presenting shallow levels only, or *"just show rule related to my own house features"* (P30), and users may query details on demand.

To carefully balance between explanation completeness and usability, if the full rules are shown, it is beneficial to **highlight *local* rule clauses** describing the current instance on top of the *global* rule explanation.

5.8 Decision Tree

5.8.1 Pros

Similar to rule, participants regarded decision tree as *"the most logical one"* (P20) that ***"tells you the decision-making process"*** (P04):

> *"(Decision tree) shows the process of thinking with AI, what it's going to do with the information."* (P10)
>
> *"How the algorithm is working, what the machine is thinking about when it's coming up with the prediction."* (P16)

5.8.2 Applicable Explanation Goals

Participants mentioned an advantage of decision tree is to **differentiate similar instances**, possibly due to its unique tree layout:

> *"It explained very well what's the difference between them (the two confusing instances)."*(P04, Bird, Report)
>
> *"It would show you how to pick up the different types of variants."* (P10, Bird, Report)
>
> *"I think this (decision tree) is the graphic comparison, like this beak might be sharper or smaller than this one, all those comparisons help"* (P09, Bird, Unexpected).

Such advantage also supports **counterfactual reasoning** by checking alternative feature values on the adjacent branches.

> *"(Decision tree) can see how to improve. It has a comparison with different outputs."* (P29, Health, Trust)
>
> *"Where does my house stands, if I'd be here, then I maybe try to change some of my features, to see how do these features affect my house price, or other houses compared to my own house."* (P30, House)

5.8.3 Cons

Several participants brought up its **weakness in communication and interpretation**.

> *"(Decision tree) is not natural language, it is more difficult to explain to my family."* (P01, House, Communication)
>
> *"This is more like a logical thing for me to see. But I wouldn't use this as an explanation to family, because that's just weird. I don't want to rack their brains too much."* (P20, House, Communication)

Indeed, in our user study, even with a two-feature two-layer decision tree, a number of participants commented:

> *"It's confusing."* (P05, Bird, Learning)
>
> *"It got too much information."* (P16, Bird, Unexpected)
>
> *"I don't really understand this one. I think it's a little bit complicated."* (P08, Bird, Learning)

Since it is less interpretable than other forms, some participants suggested to show it on-demand.

> *"I don't think these two (decision tree, decision flow chart) are necessary to show in the first UI (user interface). Maybe these two can be hidden in an icon that says 'process'. Because it (decision tree) is more like a program in process."* (P04, Bird, Trust)

Besides the tree structure, we used another flow chart visual representation (decision flow chart) in the study. In tasks that the input data were images (Bird and Car task), quite a few participants found neither the tree nor the flow chart structure helpful, and they only focused on the saliency features or objects in the flow chart.

> *"I don't think it (the flow chart structure) matters, just the head and the belly (the highlighted region shown in the flow chart) matter."* (P14, Bird)

5.8.4 Design Implications

Similar to the suggestions in rule, to reduce its complexity, one participant suggested **trimming the tree** and just showing the main branches, hiding the deeper branch and only **showing details on-demand**.

> *"You could use this one (the two-feature decision tree) as a beginning, based on this, and you click (one branch) to another in-depth version of the price calculation. Because this (price prediction) range is still very far wide, and the features given is not enough, so if you want to (check details) maybe click and (it will) add more features to it (that branch), then get a narrow range (of prediction)."* (P28, House)

Although rule-based explanations are *global* explanations (on model's overall behavior), many participants tended to focus on the branch pathway where their own input resides. By doing so, they were seeking *local* explanation (of the current input) on top of the *global* explanation. This suggests a tree-based explanation may only need to show branches containing interested instances, or **highlight branches for user's interested instances**. Users did so for the explanation goals to verify AI's decisions, and to compare with other counterfactual instances.

> *"I know there're factors that could be other houses that lead to different prices, but I still see it as, 'okay, I plug in my own numbers here and what's my price?' So it's still specific to me.'* (P20, House, Trust) – Displays local as well as global explanations
>
> *"The only thing we need is to indicate my own position on this (decision tree) branch. ....Then I can chase the features of my house."* (P30, House, Unexpected) – Suggests to highlight the pathway for user's interested instance

5.9 Input

It serves as necessary background information, and participants regarded input as a *"profile"* (P24) that *"stating the facts"* (P20). It allows participants to **understand what information AI's decision is based on**, and can help "debug" to see *"if AI is missing the most important feature"* in input (P22, Health, Bias), and *"whether or not the input is enough for it (AI) to make that decision"* (P16, Health, Trust).

When checking input, participants tended to intuitively *"look for certain features"* (P14) to judge by themselves. In the card selection & sorting, some participants used input as an anchor, put it side-by-side with example-bases explanation forms (similar, typical, and counterfactual example) for comparison. Quantitative results led to the same findings, as input was clustered together with other example-based explanation forms (Fig. 23).

5.10 Output

In our study, the output card contains prediction information of a point prediction, a prediction range, and their corresponding uncertainty level (for regression tasks); Or top three predictions and their likelihood (for classification tasks) (Fig 10). For the output information presentation, some participants preferred to check the point prediction at the beginning, and check the detailed prediction range and uncertainty level on-demand or leave them at the end, since they *"need a longer time to understand what these numbers mean"* (P02).

Participants had divergent preferences and understandings on the prediction presentation forms. Compared to a point prediction (e.g.: the prediction on house price is 650k), some preferred to see a **prediction range** in regression tasks (e.g.: house price is 638-662k), or top predictions list in classification tasks, because such prediction range *"give choices"* (P05, Bird classification task, Differentiation), *"acknowledges a possibility"* (P18, Bird, Unexpected), rank the decision priorities (P03, Car classification task, Safety), help them *"(the range) to see how different between my and AI prediction"* (P01, House regression task, Unexpected), and provide rooms for adjustment and negotiation:

> *"If I want to sell it higher, and I'll put 662k (the upper bound). Or if I wanted to sell it fast, then I'll put 638k (the lower bound). There's always a range, it's not necessarily just one price. And people will always bargain too."* (P20, House, Communication)

And sometimes they *"don't even need to know the (prediction) number exactly. This (range) tells me that (my diabetes risk) it's high. I have to do something. So that's what I want to know"* (P17, Health regression task, Trust), and the range gives a higher certainty than a single point prediction which enhanced participants' trust.

In contrast, some other participants were more acceptable to a narrower range or a **point prediction**, because they saw a wider range of prediction had its drawbacks: *"(the prediction range) shows too much fluctuation"* (P07, House, Trust); And seeing the full predictions list (some with lower prediction likelihoods) may make them confused and discredit AI's decisions. Thus a narrower range may give them more confidence about AI's prediction.

> *"Seeing that the range is pretty small makes me a lot more confident that they've got enough data to actually be drawing conclusions."* (P16, Health, regression task, Trust)

For the prediction likelihood/uncertainty/confidence[2], some participants **had a hard time understanding the meaning of uncertainty** and required researcher's further explanations. A high certainty *"reassure AI's performance"* (P22), *"help a lot of persuading yourself into believing in AI"* (P10), which is consistent with the recent quantitative finding on certainty level and trust calibration [24]. Especially when AI and users disagree with each other (unexpected), participants may abandon their own judgment due to AI's high certainty.

> *"If it had a high certainty, then I would want to know why my estimation is wrong."* (P10, House, Unexpected)

[2]Although the AI community has distinct methods to compute output likelihood and uncertainty level, in our study we used likelihood, confidence and uncertainty interchangeably to avoid participants' confusion.

5.11 Performance

After checking the performance information, most participants realized the probabilistic nature of AI decisions: *"AI is not perfect"* (P20), *"they (AI) make errors sometimes"* (P05). If the performance is within their acceptable range, participants would accept the "imperfect AI". And seeing the performance level helped users to **set a proper expectation for AI's performance**.

> *"I get it's downside. Performance warns me to, 'Hey, you know, it's not really accurate. There's some room for error.' "* (P24)

And sometimes participants may calibrate their trust according to the error rate (in classification tasks) or error margin (in regression tasks).

> *"If there is a really big margin (of error), then it would probably demean the trust."* (P23)

Almost all participants understood the meaning of accuracy (error rate) in classification tasks, whereas many participants had a difficult time understanding the margin of error in regression tasks.

> *"Performance is really in detail. I mean not everyone is familiar with statistics, like mean error."* (P30, House, regression task)

Unlike the uncertainty level in output (Section 5.10) which is case-specific decision quality information, a few participants noticed performance is model-wide information, and just provides *"general information showing the trust level of the system"* (P04) is *"too general, I would want to know specifically why (the speed) it's going down in this particular case of driving"* (P05, Car, Unexpected). Thus, they suggested there was no need to show it every time, *"you should know before you use AI"* (P11).

However, in some particular explanation goals such as to detect bias, participants may require to check the fine-grained performance analysis on interested outcome.

> *"It (fine-grained performance on road/weather conditions) explains how often I should be confident in rainy days."* (P19, Car, Safety)

5.12 Dataset

In our study, the dataset card contains training dataset distribution of the prediction outcomes. Even after researchers' explanation, some participants did not well understand or **misunderstood** the information on this card (for example, some misinterpreted the distribution graph as feature shape), indicating it requires a higher level of AI/math/visualization literacy [13, 3, 8]. For those who comprehended the dataset information, some participants tended to link the dataset size with model accuracy and trust.

> *"The higher the (training data distribution) curve goes, then I would be more confident that they have a big pool of data to pull from."*(P31, Health, Unexpected)

Some participants intuitively wanted to check their own data point within the training data distribution, and use it as a dashboard to **navigate, identify, and filter interested instances** (such as similar, typical, and counterfactual examples), to compare what are the same and different features between their input and the interested instances.

> *"I want to see which region I fall in the population, and compare with people around to see why my (diabetes) risk is only 10% with a family history."* (P01, Health, Unexpected)

Nevertheless, in practice there may be some restrictions on reviewing the detailed dataset information due to data proprietary and privacy, as brought out by P19:

> *"I want to know the number of data and the details of it to verify. But I don't know if that's going to be able to be viewed. That's probably secret, right?"* (P19, House, Expected)

6 Participants' Information

Participant number	Age	Sex	Education level	Major or Industry	AI familiarity	AI attitudes	Tasks	Interview duration (min)
P01	38	M	Bachelor	computer science	program but not in AI	interested	House; Health; Car; Bird	120
P02	26	M	PhD	HCI	program but not in AI	concerned; interested; excited	Health; Bird	90
P03	29	F	PhD	HCI	use AI (Google) to reminders/navigation/daily use/play music or video etc.	interested	House; Car	74
P04	28	M	Master	HCI	program but not in AI	concerned; interested; excited	House; Car	94
P05	40	F	Trade	editing	heard	concerned; interested; excited	Car; Bird	46
P06	21	F	Some college credit	psychology	use AI (Google home) to play music	concerned; skeptical; interested	Health; Bird	76
P07	62	M	Bachelor				House; Car	64
P08	22	F	High school	computer science	program but not write AI code	excited	Health; Bird	55
P09	40	M	Bachelor	Business development and sales (IT)	use AI (Google navigator) to traffic and directions	excited	Car; Bird	51
P10	19	M	High school	cooking	heard	neutral	House; Bird	54
P11	30	F	Bachelor	IT	program but not write AI code	interested	House; Bird	76
P12	48	F	High school		heard	neutral	House; Bird	74
P13	53	M	Bachelor	customer service	heard	concerned; skeptical	Health; Car	69
P14	47	M	Some college credit	healthcare-sterilization work	never	interested	House; Bird	55
P15	73	M	Professional	retired	heard	skeptical	Car	81

Participant number	Age	Sex	Education level	Major or Industry	AI familiarity	AI attitudes	Tasks	Interview duration (min)
P16	34	F	Professional	law	heard	concerned; interested; excited	Health	67
P17	70	M	Bachelor	retired	heard	neutral	Health; Car	47
P18	27	M	Some college credit	General studies and legal studies	heard	skeptical; neutral; excited	Bird	41
P19	35	F	Bachelor	Government or social services (employment services for indigenous peoples)	heard	concerned; skeptical; interested; excited	House; Car	42
P20	30	M	Bachelor	Food industry	heard	concerned; skeptical; interested; excited	House	58
P21	26	F	Bachelor	Interior designer	use AI (chatting with clients)	concerned; interested; excited	Car	60
P22	23	F	Some college credit	Student (RMT); Work (hospitality (restaurant)	heard	concerned; skeptical; excited	Health	69
P23	31	M	Master	Accountant	use AI (google Home) to preferred music/movie	excited	Health	72
P24	41	M	Bachelor	Financial industry	use AI (investment software) to help drive investment decisions	excited	Health	69
P25	72	M	Master	retired	heard	concerned; interested; excited	Health	112
P26	70	F	Bachelor	retired	heard	skeptical; interested	Bird	52
P27	28	F	Bachelor	hospitality	heard	interested	Car	45
P28	28	M	Trade	Marlcotins sale	heard	interested	Health	88

Participant number	Age	Sex	Education level	Major or Industry	AI familiarity	AI attitudes	Tasks	Interview duration (min)
P29	43	F	Bachelor	Project management in construction (currently no job)	heard	concerned; interested; excited	House	67
P30	24	F	Master	Computer science	program but not write AI code	concerned	House	83
P31	25	F	Bachelor	psychology office worker	heard	interested	Health	65
P32	39	F	Bachelor	car insurance	heard	excited	Car	59

7 Study Material

7.1 Demographic Questionnaire

1. Your age: _____
☐ prefer not to disclose
2. Your gender:
☐ Female
☐ Male
☐ Other
3. What is the highest degree or level of school you have completed or currently enrolled?
☐ No schooling completed
☐ Nursery school to 8th grade
☐ Some high school, no diploma
☐ High school graduate, diploma or the equivalent (for example: GED)
☐ Some college credit, no degree
☐ Trade/technical/vocational training
☐ Bachelor's degree
☐ Master's degree
☐ Professional degree (e.g. MD, JD)
☐ Doctorate degree (PhD)
4. If you are a student, what is your major? If you are working, what is your current work industry?
5. What is your understanding of artificial intelligence (AI)?
☐ I have never heard of AI before
☐ I only hear of AI from the news, friends, etc.
☐ I use AI in my work or life. If so, please specify what kind of AI do you use: _____ , to accomplish what tasks:_____
☐ I can program, but I can not write AI code
☐ I can write AI code
6. What is your opinion on incorporating AI technology into our everyday decision-making scenarios? (you can select multiple choices)
☐ I am not interested in AI, and I do not pay attention to it
☐ I am concerned about the prevalence of AI (e.g.: it will take over many people's job; it's a threat to human beings)

☐ I am skeptical of the incorporation of AI technology, but I would like to learn more about it
☐ I am neutral regarding the incorporation of AI technology
☐ I am interested in the incorporation of AI, and willing to know more about it
☐ I am excited to use AI to improve my work and life

7.2 Interview Material

We attach the interview material used in the study at the end of the Appendix, including:

1. The four tasks shown as storyboards;
2. The explanation goals shown as storyboards;
3. The explanation forms shown as cards.

References

[1] Diabetes Prediction Dataset, 2020. Accessed: 2020-09-10.

[2] The Boston Housing Dataset, 2020. Accessed: 2020-09-10.

[3] Jeremy Boy, Ronald A. Rensink, Enrico Bertini, and Jean-Daniel Fekete. A principled way of assessing visualization literacy. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1963–1972, 2014.

[4] Virginia Braun and Victoria Clarke. Thematic analysis. In *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological.*, APA handbooks in psychology®., pages 57–71. American Psychological Association, Washington, DC, US, 2012.

[5] Andrea Bunt, Matthew Lount, and Catherine Lauzon. Are explanations always important? In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces - IUI '12*, page 169, New York, New York, USA, 2012. ACM Press.

[6] Federica Di Castro and Enrico Bertini. Surrogate Decision Tree Visualization, 2019.

[7] Finale Doshi-Velez and Been Kim. Towards A Rigorous Science of Interpretable Machine Learning. feb 2017.

[8] Iddo Gal. Adults' statistical literacy: Meanings, components, responsibilities. *International Statistical Review*, 70(1):1–25, 2002.

[9] Shirley Gregor and Izak Benbasat. Explanations from Intelligent Systems: Theoretical Foundations and Implications for Practice. *MIS Quarterly*, 23(4):497, dec 1999.

[10] Robert R. Hoffman and Gary Klein. Explaining explanation, part 1: Theoretical foundations. *IEEE Intelligent Systems*, 32(3):68–73, 2017.

[11] Brian Y. Lim and Anind K. Dey. Assessing demand for intelligibility in context-aware applications. In *Proceedings of the 11th international conference on Ubiquitous computing - Ubicomp '09*, page 195, New York, New York, USA, 2009. ACM Press.

[12] Brian Y Lim, Qian Yang, Ashraf Abdul, and Danding Wang. Why these Explanations? Selecting Intelligibility Types for Explanation Goals. page 7, 2019.

[13] Duri Long and Brian Magerko. What is ai literacy? competencies and design considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–16, New York, NY, USA, 2020. Association for Computing Machinery.

[14] Helena Matute, Fernando Blanco, Ion Yarritu, Marcos Díaz-Lago, Miguel A. Vadillo, and Itxaso Barberia. Illusions of causality: how they bias our everyday thinking and how they could be reduced. *Frontiers in Psychology*, 6(July):1–14, 2015.

[15] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A Survey on Bias and Fairness in Machine Learning. 2019.

[16] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

[17] Yao Ming, Huamin Qu, and Enrico Bertini. RuleMatrix: Visualizing and Understanding Classifiers with Rules. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):342–352, jan 2019.

[18] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation. feb 2018.

[19] Ingrid Nunes and Dietmar Jannach. A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction*, 27(3-5):393–444, 2017.

[20] Mireia Ribera and Agata Lapedriza. Can we do better explanations? A proposal of user-centered explainable AI. In *Joint Proceedings of the ACM IUI 2019 Workshops*, 2019.

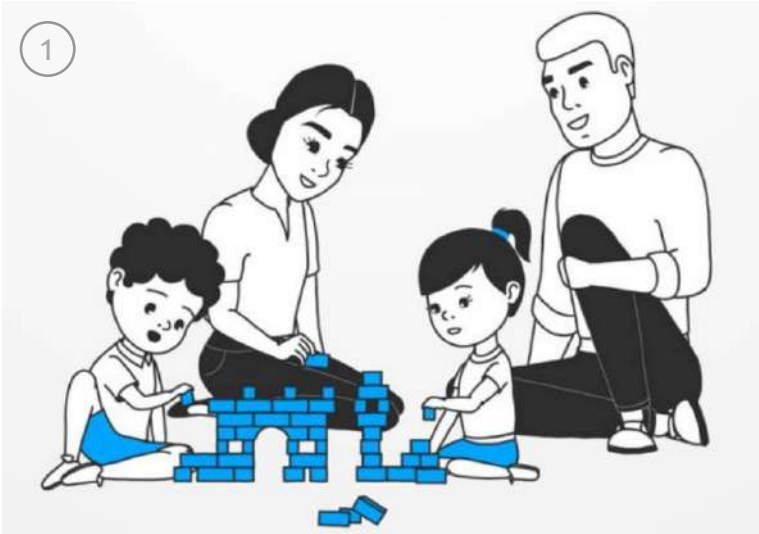
[21] Amy Turner, Meena Kaushik, Mu-Ti Huang, and Srikar Varanasi. Calibrating Trust in AI-Assisted Decision Making. 2020.

[22] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

[23] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving video database with scalable annotation tooling. *CoRR*, abs/1805.04687, 2018.

[24] Yunfeng Zhang, Q. Vera Liao, and Rachel K.E. Bellamy. Efect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 295–305, 2020.

Selling your house

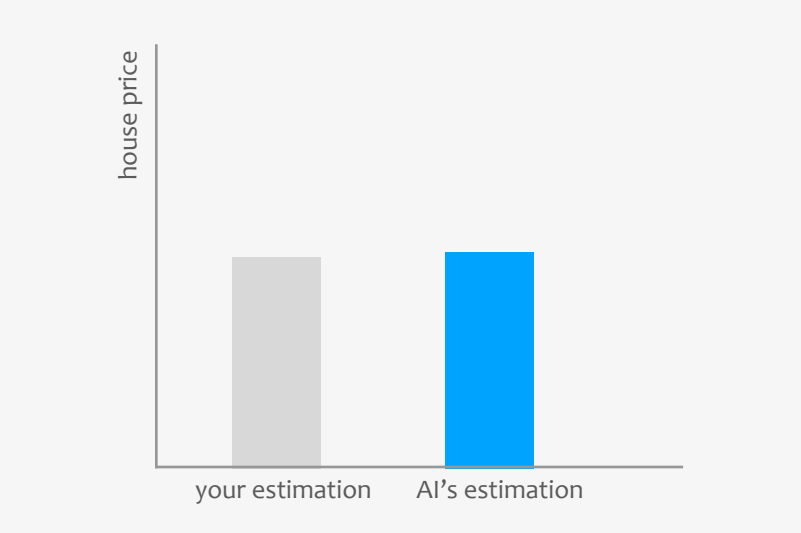


Suppose your family is expanding

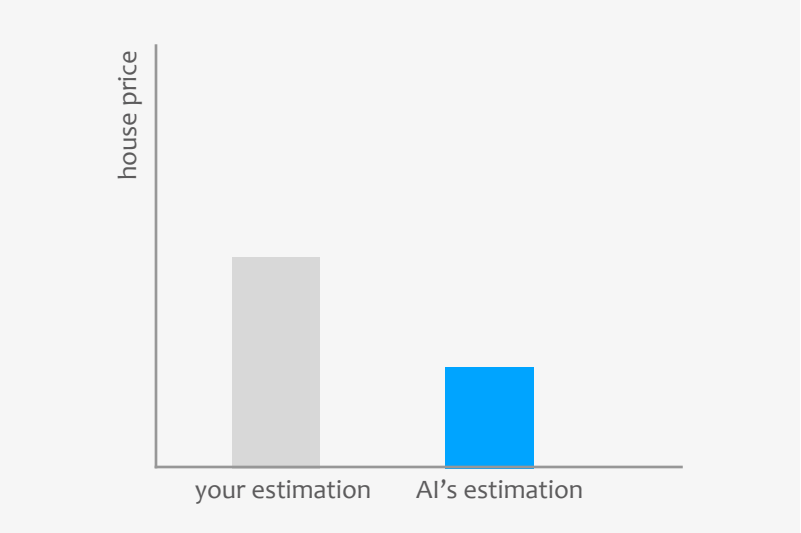
and you need to sell your current house, for a bigger one.

Since your budget is limited, you need to sell your current house at a really good price

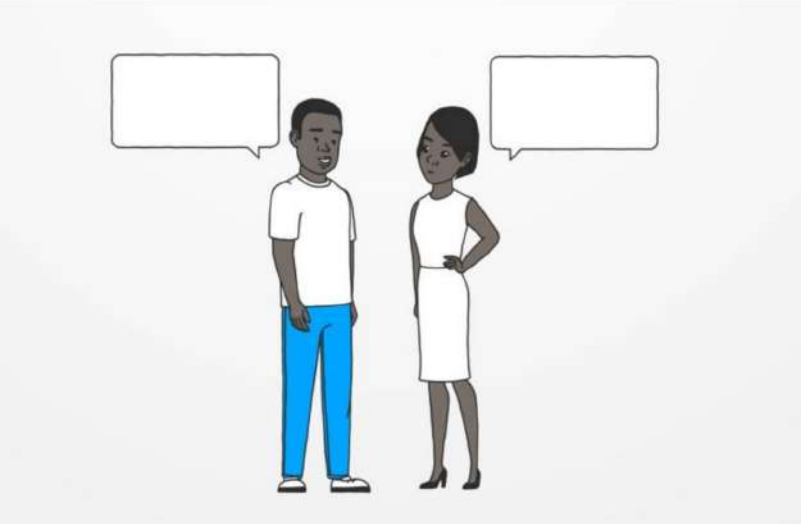
You get to know there is an **aritifical intelligence (AI)** tool that can **predict house price**. It may help you to get a propoer estimate of your house.



AI's prediction ***aligns*** with your own estimation



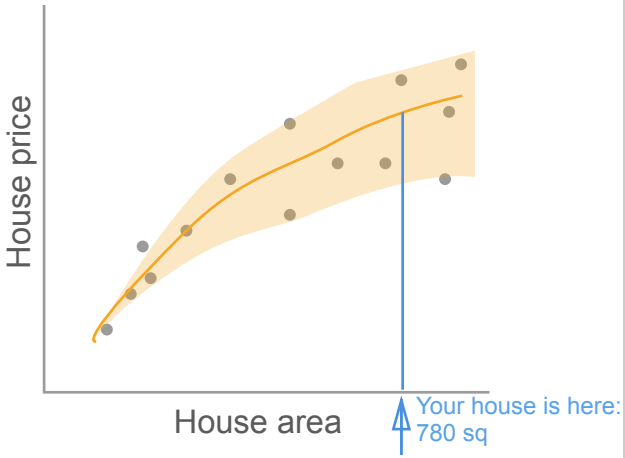
AI's prediction does ***not align*** with your estimation

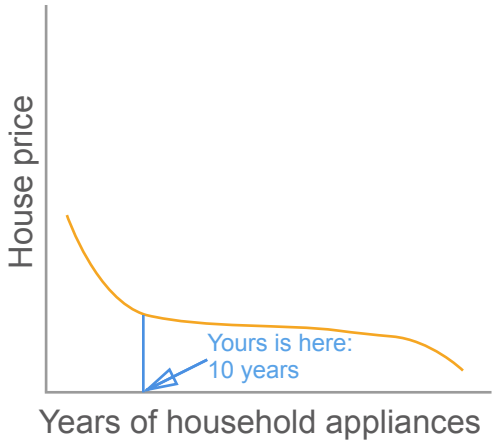


You need to ***communicate*** your decision with your family

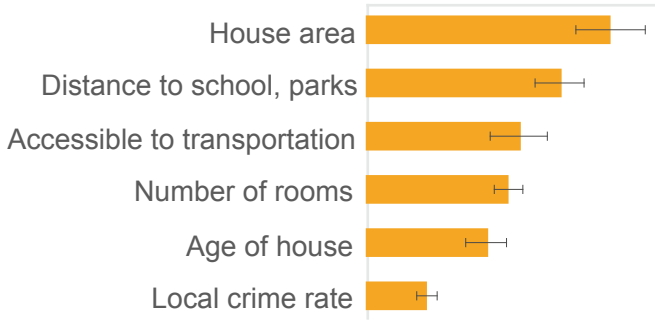
You doubt whether to ***trust*** the AI tool or not

You need to decide whether to do a renovation or replacement of appliances to increase your house value, and ***which action*** is the most cost-effective.

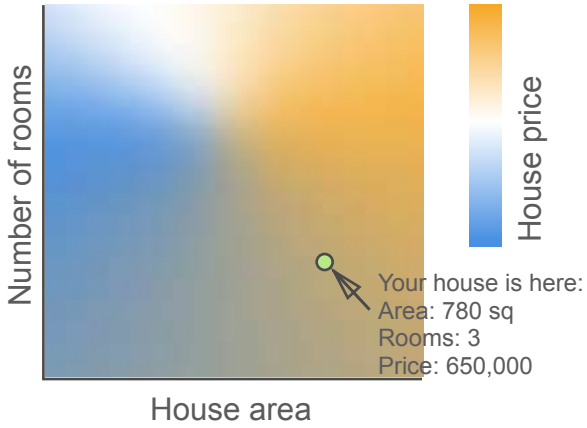




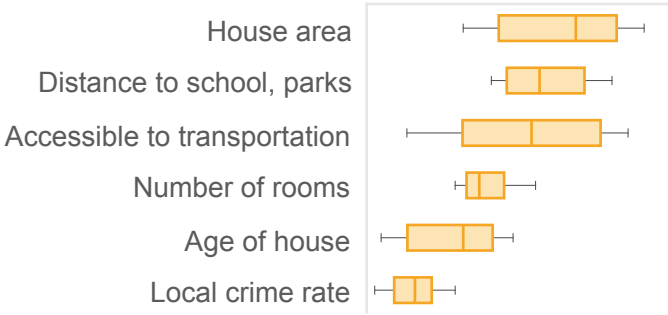
How important is each feature to the result:



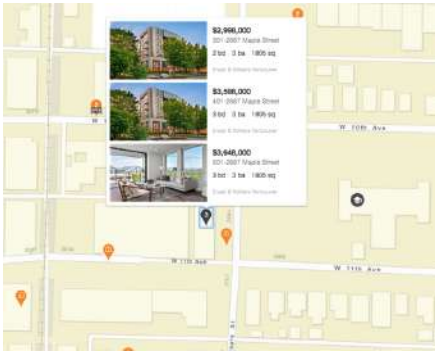
The houses of **similar *price*** as yours



Feature importance score



The houses of **similar *features*** as yours



A **typical** house to sell at the estimated price as yours is like:

In your neighbourhood:

- 2 bedrooms
- 2 bathrooms
- 1000 sq
- 20 years old
- ……

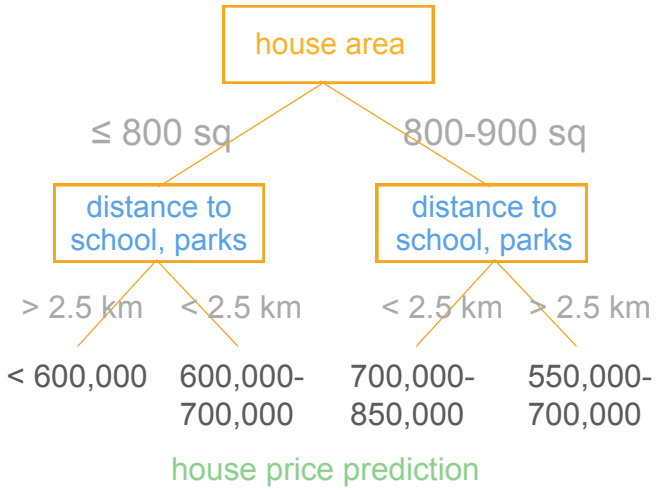
If the feature of your house had changed to the following feature,
your house price would have increased by 10%:

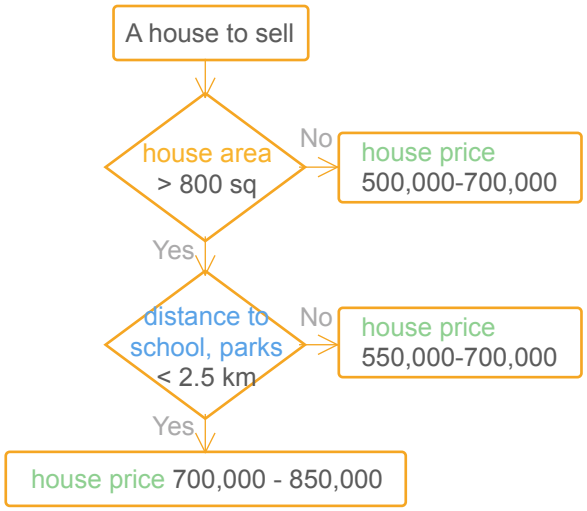
- have a back yard, or
- 3 bathrooms, or
- 1200 sq, or
- less than 10 years old, or
- has new household appliances
- ……

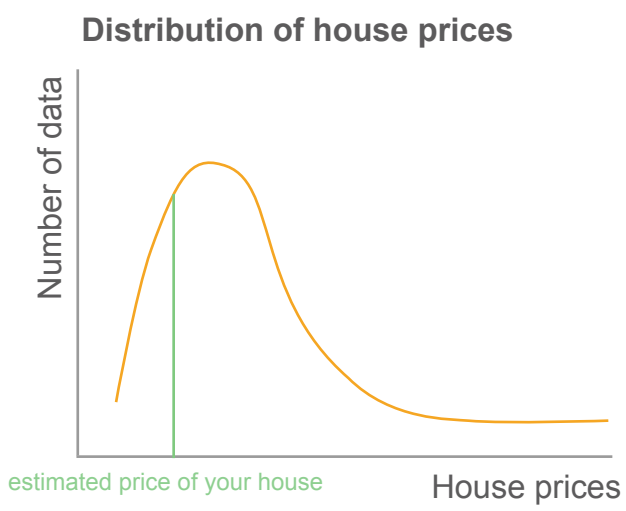
If **house area** ≤ 800 sq,
and **distance to school, parks** > 2.5 km,
Then house price **is no more than 600,000**

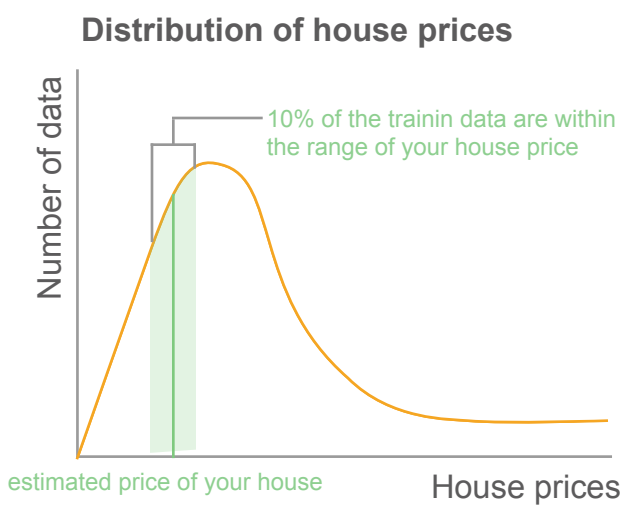
If **house area** is 800 - 900 sq,
and **distance to school, parks** < 2.5 km,
Then house price **is about 700,000-850,000**

	house area	distance to school, parks	house price prediction
Rule 1	≤ 800 sq	> 2.5 km	< 600,000
Rule 2	≤ 800 sq	< 2.5 km	600,000-700,000
Rule 3	800-900 sq	< 2.5 km	700,000-850,000









The features of your own house

- 2 bedrooms
- 1 bathroom
- 780 sq
- 20 years old
- household appliances for 10 years
- distance to school, parks: 2 km

Predicted price of your own house

$ 650,000

Predicted price of your own house

$ 650,000

with certainty of 90%

$ 638 ~ 662,000

with certainty of 95%

The performance of the AI house precdiction tool

- Mean prediction error: ± 50,000
- Max prediction error: ± 120,000
- The AI tool can explain 95% of the variation in the training data

Personal health decision

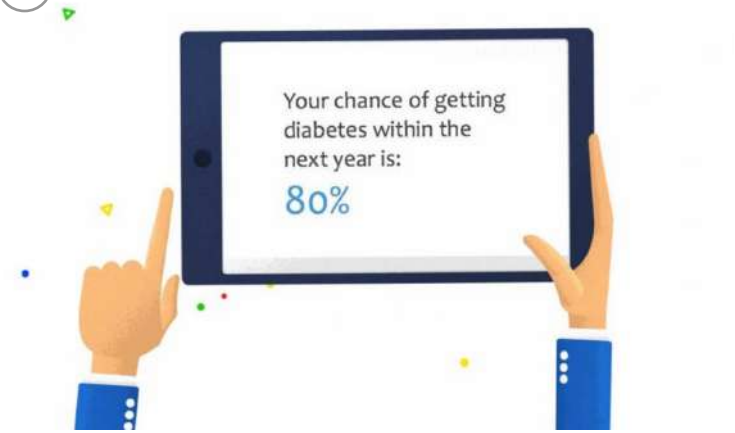
1

Suppose one day you received an email from the company that stores your health record,

2

and it can provide you a service that help you identify your risk of diabetes, by analyzing your health records.

3



You gave it a try, and it tells you that you have 80% percent of chance to be diagnosed with diabetes in the next year.

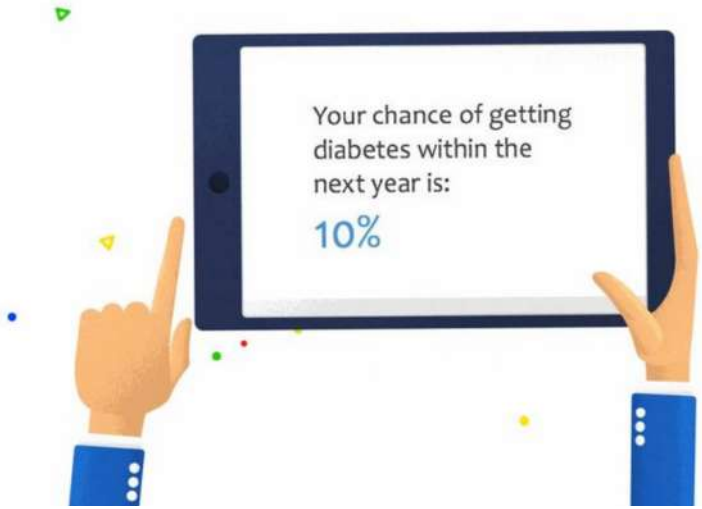
4

Would you like to take the predicted result from the software?

You doubt whether to ***trust*** the software prediction on your diabetes risk.

You want to know how to ***adjust your lifestyle*** accordingly to lower the risk of diabetes.

You need to need to inform ***family*** members and consult your ***doctor***.



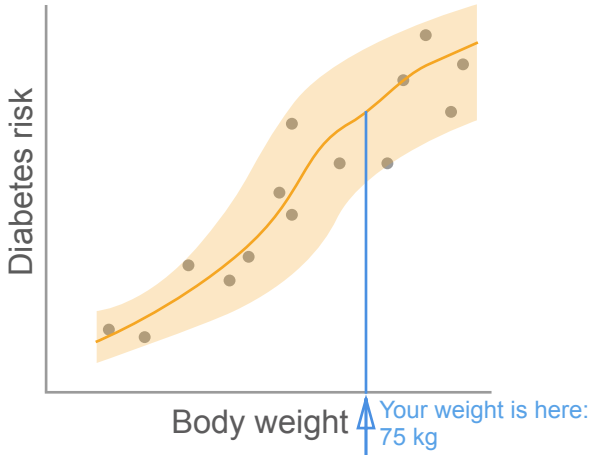
It predicts your chance of getting diabetes is ***low***.

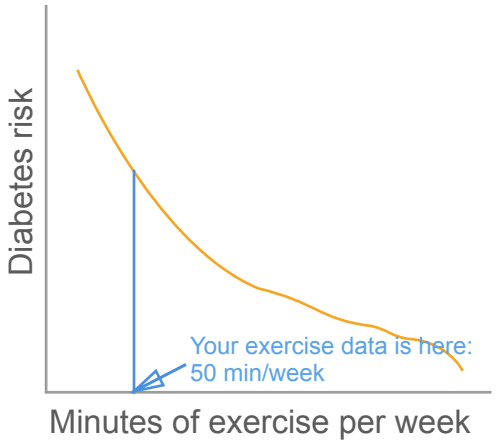
Diabetes tends to run in your family, and you're afraid of getting it someday.

You maintain ***good health*** with no major diseases or a family history of diabetes.

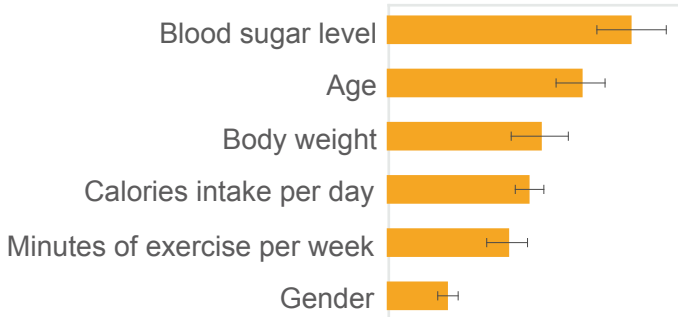
You're aware that the insurance company may use such a prediction from the software to ***determine your insurance fee and benefits***.

You doubt whether the software will perform the same among people with ***different gender, age, or ethnicity group***.



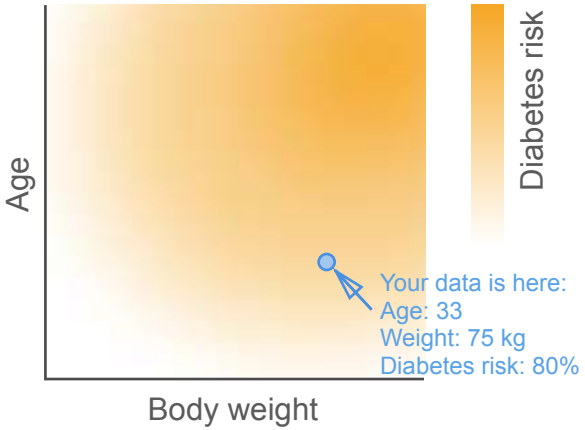


How important is each feature to the result:

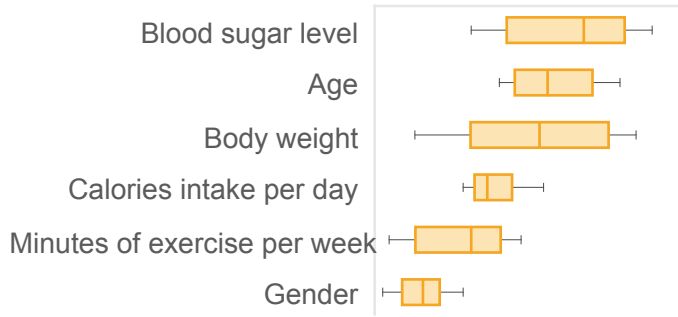


The case that has the **similar *diabetes risk*** as yours:

- Male, 32 years old
- Three consecutive blood sugar level: higher than normal, higher than normal, normal
- Body weight: 80 kg, height 178 cm
- Calories intake per day: 2900
- Minutes of exercise per week: 30 min
- Family history of diabetes: ……
- .....



Feature importance score



The case that has **similar *features*** as yours:

- Male, 35 years old
- Three consecutive blood sugar level: normal, normal, higher than normal
- Body weight: 81 kg, height 183 cm
- Calories intake per day: 3400
- Minutes of exercise per week: 60 min
- Family history of diabetes: ……
- .....

A **typical** case of the same diabetes risk as yours is like:

- Male, 45 years old
- Three consecutive blood sugar level: normal, normal, higher than normal
- Body weight: 78 kg, height 175 cm
- Calories intake per day: 3000
- Minutes of exercise per week: 30 min
- Family history of diabetes: ……
- .....

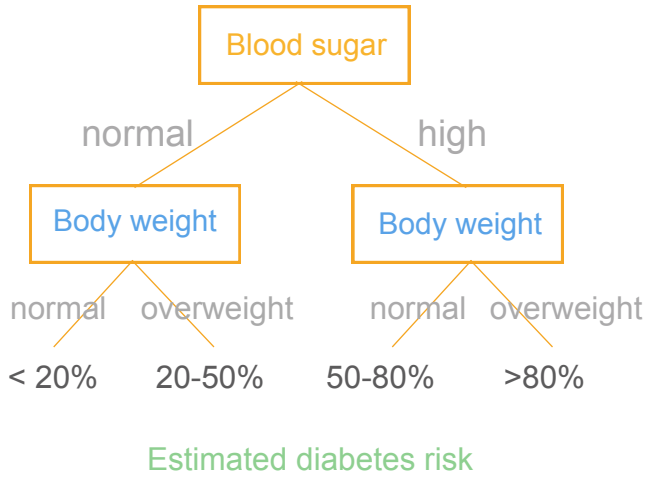
If your health data had changed to the following, your diabetes risk would have decreased by 20%:

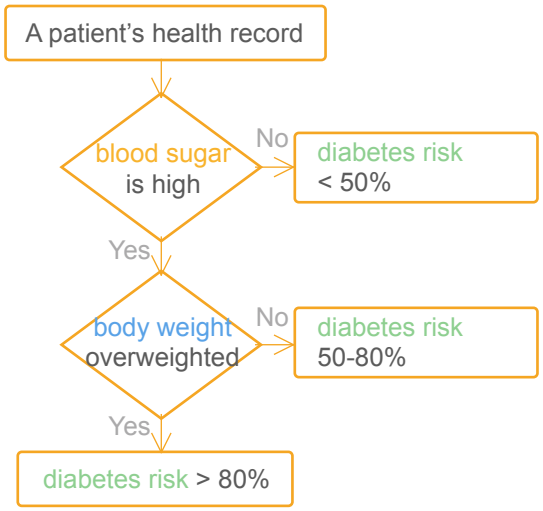
- 3 years younger than now
- Body weight: loss 5 kg
- Increase 50 min of weekly exercise
- Reduce 500 calories of daily calories intake
- ……

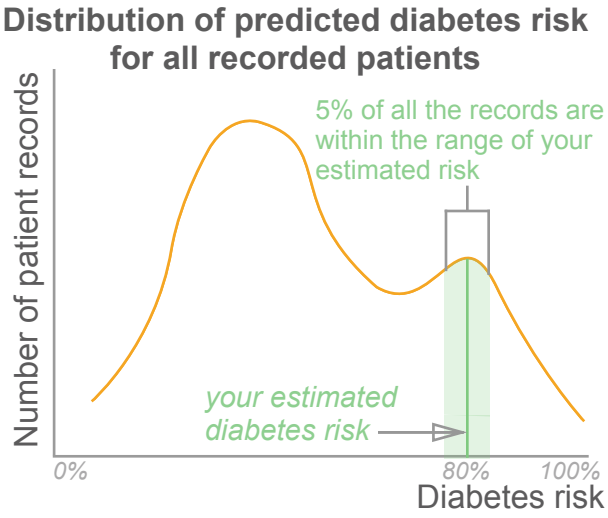
If **blood sugar** is **high**,
and **body weight** is **overweighted**,
Then the estimated diabetes risk
is above 80%

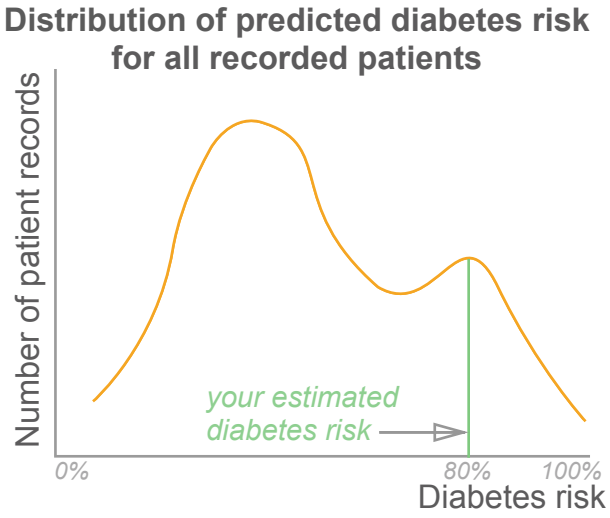
If **blood sugar** is **normal**,
and **body weight** is **overweighted**,
Then the estimated diabetes risk
is about 20-50%

	blood sugar	body weight	diabetes risk
Rule 1	high	high	> 80%
Rule 2	high	normal	50-80%
Rule 3	normal	normal	< 20%









The data from your health records used for prediction:

- Male, 33 years old
- Three consecutive blood sugar level: normal, normal, higher than normal
- Body weight: 75 kg, height 175 cm
- Calories intake per day: 3200
- Minutes of exercise per week: 50 min
- Family history of diabetes: ……
- .....

Your chance of getting diabetes within the next year is:

80 %

Your chance of getting diabetes within the next year is:

80 %

with a certainty of 90%

75 ~ 85%

with a certainty of 95%

The performance of the AI tool to predict diabetes risk

- Mean prediction error: ± 15%
- Max prediction error: ± 30%
- The AI tool can explain 75% of the variation in the training data

Buying an autonomous driving vehicle

1

You're test-driving an autonomous driving vehicle

2

Equipped with sensors and artificial intelligence (AI) system, the car can drive on its own.

3

Your main concern is the safety issue.

4

You need to decide whether to buy the car or not.

You notice the car sometimes drives much ***slower than the expected speed limit.***

You're easy to get motion sickness, and you notice you seem to get ***car sick*** more frequently ***in autopilot mode.***

You need to ***communicate*** with your family about your judgement on the car's safety.

You want to know if the autopilot mode performs equally ***under different road, weather*** conditions, and during the ***night.***

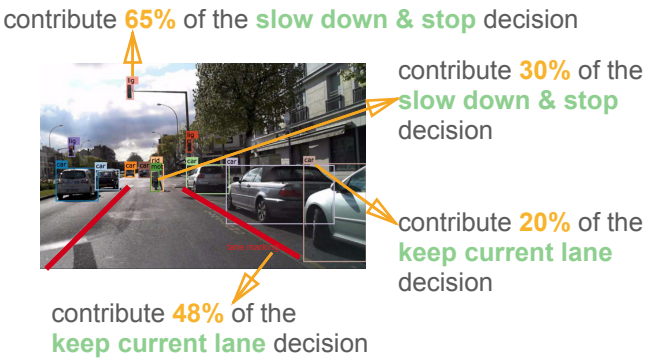
You need to know whether the autopilot mode is ***safe*** and ***reliable.***

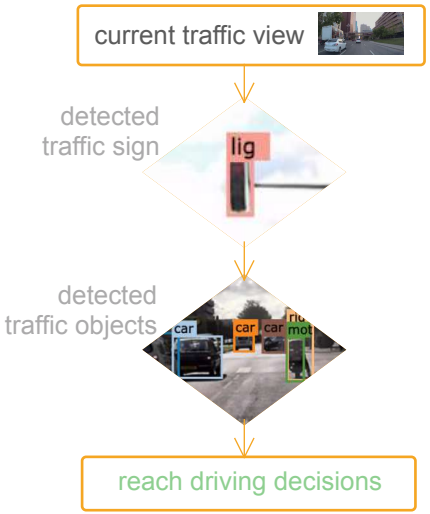
Similar traffic conditions as the current one, from the dataset to train the self-driving car:

Typical traffic conditions to reach the self-driving car's current decison:

Keep straight 95%	**Keep current speed** 70%	**Right lane change** 55%

Important objects detected for the self-driving car's judgement:

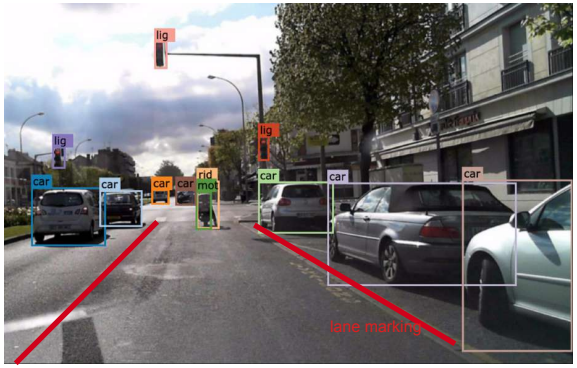




The current driving decisions, and **their percentage in the training dataset** where the self-driving car learns from

	Confidence	*Precentage*
Keep straight	95%	25%
Keep current speed	95%	34%
Right lane change	55%	2.9%

Important objects detected for the self-driving car's judgement:



Current traffic view:

Driving decisions under the current traffic:

	Confidence
Keep straight	95%
Keep speed at 50 km/h	95%
Right lane change	55%

Overall performance of the autonomous driving mode:

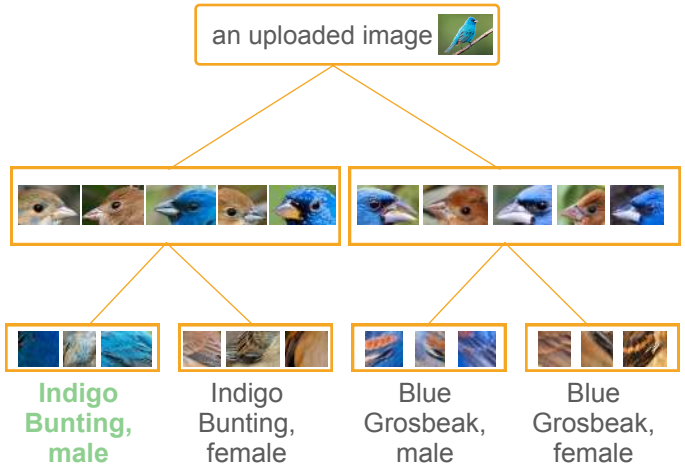
*Measured using average distance driven between disengagements**

- Under **normal** road condition: 40 km
- During the **night**: 5 km
- On **rainy** days: 3 km
- On **snowy** days: 1 km

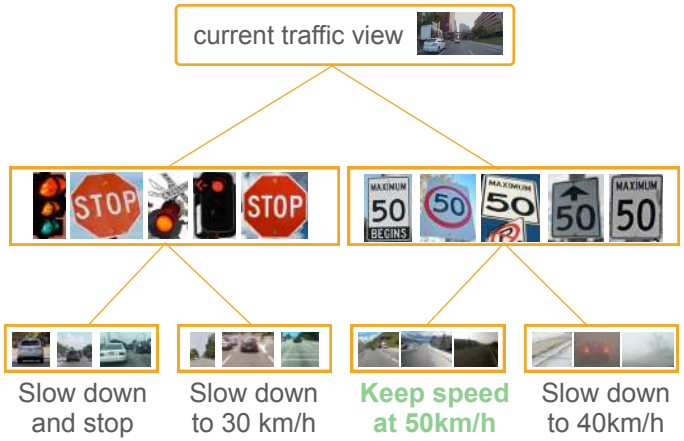
* Disengagement means when the automated system is switched off by the intervention of a human driver

If **bird bill** is **small and thin**,
and **wings and tails** are **short**,
Then the bird is recognized as
Indigo Bunting

If **bird bill** is **big and thick**,
and **wings and tails** are **long**,
Then the bird is recognized as
Blue Grosbeaks



Bird A **>> progressive transition >>** **Bird B**



Bird A **highlight different regions** **Bird B**

If **traffic sign** is **stop sign**,
or the speed of the **car in front** are
slower,
Then the speed decision is to
slow down and stop

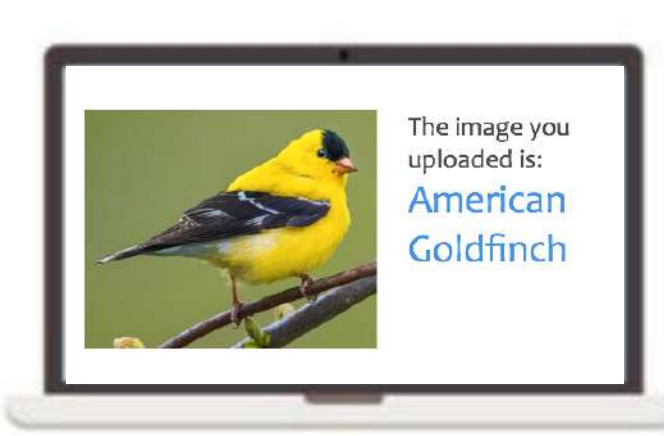
If **traffic sign** is **50km/h speed limit**,
and the speed of the **car in front** are
the **same or faster**,
Then the speed is kept at
50km/h

Learning bird species

Suppose you're a biology student, and are studying over the weekend to prepare for exam on bird species.

You get to know a bird taxonomy website that can automatically recognize the bird images you upload.

So you give it a try by uploading a bird image, and it gives you the most likely bird species.



Will you use the website to help you prepare for the exam?

You don't know whether to ***trust*** the results from the website or not.

The results sometimes does ***not align*** with your knowledge.

In the exam, you need to **write a short statement** on how you **recognize** the bird as such species.

In the exam, you need to **write *a short statement*** to ***differentiate different birds.***

Is it a good tool to improve your ***learning*** and help you know more about bird taxonomy?

Similar images to the one you uploaded:


Indigo Bunting
95%


Indigo Bunting
95%


Blue Grosbeak
70%


Blue Grosbeak
70%


Lazuli Bunting
55%


Painted Bunting
45%

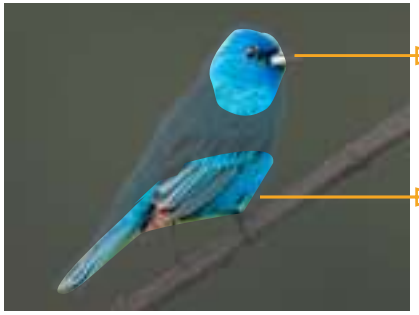
The three most likely bird according to your uploaded image, and **typical examples**

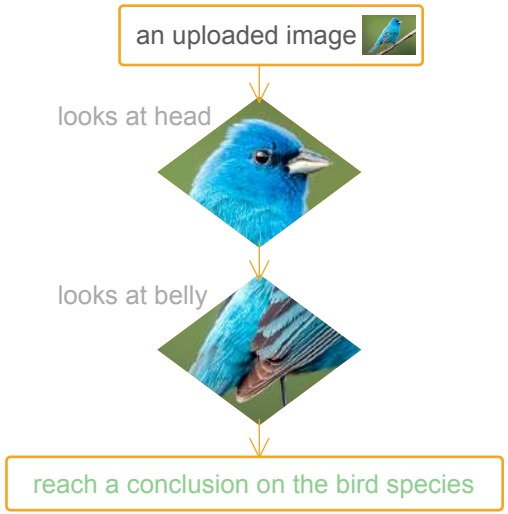
Indigo Bunting
95%

Blue Grosbeak
70%

Lazuli Bunting
55%

Important regions (highlighted) for AI's bird recognization:





The three most likely bird according to your uploaded image, and **their percentage in the training dataset** where the AI learns from

	Likelihood	*Precentage*
Indigo Bunting	95%	1.5%
Blue Grosbeak	70%	1.2%
Lazuli Bunting	55%	1.3%

Important regions (highlighted) for AI's bird recognization:

The image you uploaded:

The image you uploaded is recognized as:

	Likelihood
Indigo Bunting	95%
Blue Grosbeak	70%
Lazuli Bunting	55%

Overall performance of the AI bird recognition tool:

- Accuracy: 85%
- Error rate: 15%