EUCA: the End-User-Centered Explainable AI Prototyping Framework
Supplementary Material S2

1 INTERVIEW SCHEDULE - ADDITIONAL INFORMATION

1.1 Round 1 (RQ2): End-Users' General Requirements for Various Explanation Goals

We began the user study by introducing the researchers, 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. 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. Figure 1 shows an example 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.

Fig. 1. The interview storyboard of the Health Task in the user study

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

Author's address:

was shown as a storyboard picture, and the researcher read the text on the picture to introduce the explanation goal. Figure 2 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.

In this process, we aimed to understand user's requirements under different explanation goals before showing them the explanatory form prototyping cards.

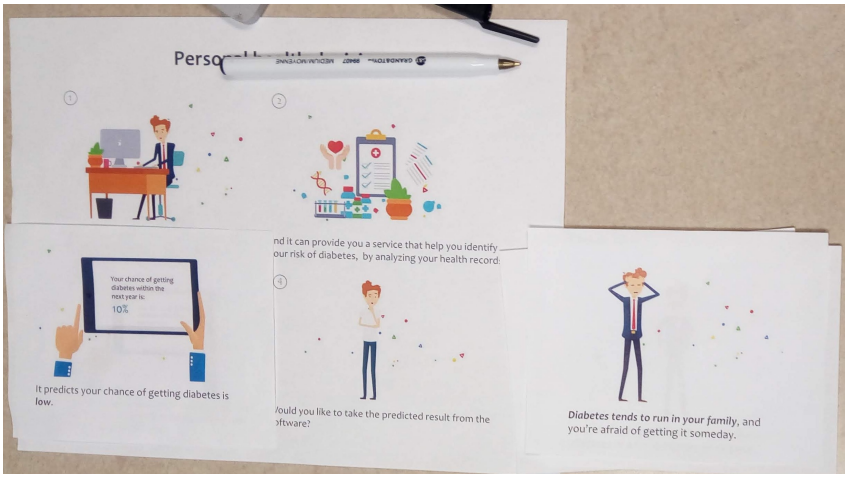


Fig. 2. The explanation goal of unexpected in the Health task context. The end-user may expect to have a high risk of diabetes due to family history. However, AI predicts the risk is only 10% which may not align with the user's expectation.

1.2 Round 2 (RQ2): Card Selection & Sorting

After discussing all the explanation goals for one task, the participant entered the second round. They again talked about their comments and requirements for each explanation goal *using* the explanatory form prototyping cards, and had a card selection & sorting.

The participants first revisited the task. Then the researcher walked through the created prototyping cards showing the explanatory forms for that task. 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 explainability needs. The participants could comment on any card anytime during this process. They could also sketch on blank cards to create new prototyping cards, and add the newly created 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 explainability needs.

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 4.1). The study session duration is 67.9 ± 18.8 (Mean ± SD) minutes (Median: 67 min, Range: 41 - 120 min). All study materials including the storyboards of tasks and explanation goals, and prototyping cards of explanatory forms are listed at the end of this document.

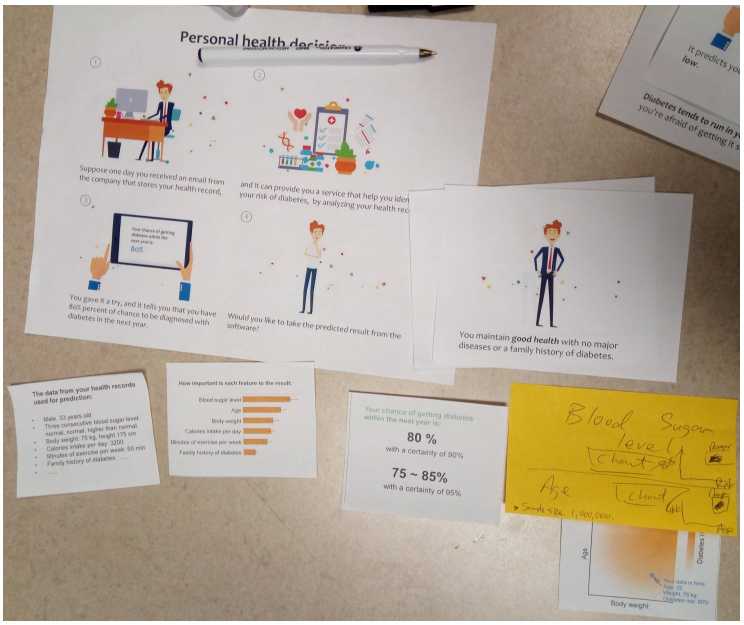


Fig. 3. **The card selection & sorting result from one participant**. Given the task and explanation goal, the participant selected and sorted the prototyping cards from left to right according to their usefulness. She also sketched to improve the last card on feature interaction.

2 QUANTITATIVE DATA ANALYSIS AND RESULTS

2.1 Quantitative Data Analysis Methods

We performed a Pearson's Chi-squared test to discover if there is an interrelation between demographic factors and the following dependent variables: participants' response to "Accept AI", "Require XAI", whether their explainability "Needs fulfilled" by the explanatory form combination, and the selection of explanatory forms for each explanation goal.

To test whether the sorting of explanatory 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 explanatory 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 explanatory 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 explanatory 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 explanatory forms into

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

a 2-dimensional space using multidimensional scaling (MDS) and visualized it. We then clustered the explanatory 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.

2.2 Do end-users need AI and explanations?

To investigate if end-users need AI and require explanations under a variety of explanation goals, we recorded participants' answers to the following two questions in the first round of the interview: "do you want to use AI as an assistant in this task and for the current explanation goal?", and "do you require additional information/explanation from AI?". A total of 300 and 293 effective responses were collected for the two questions respectively from 32 participants.

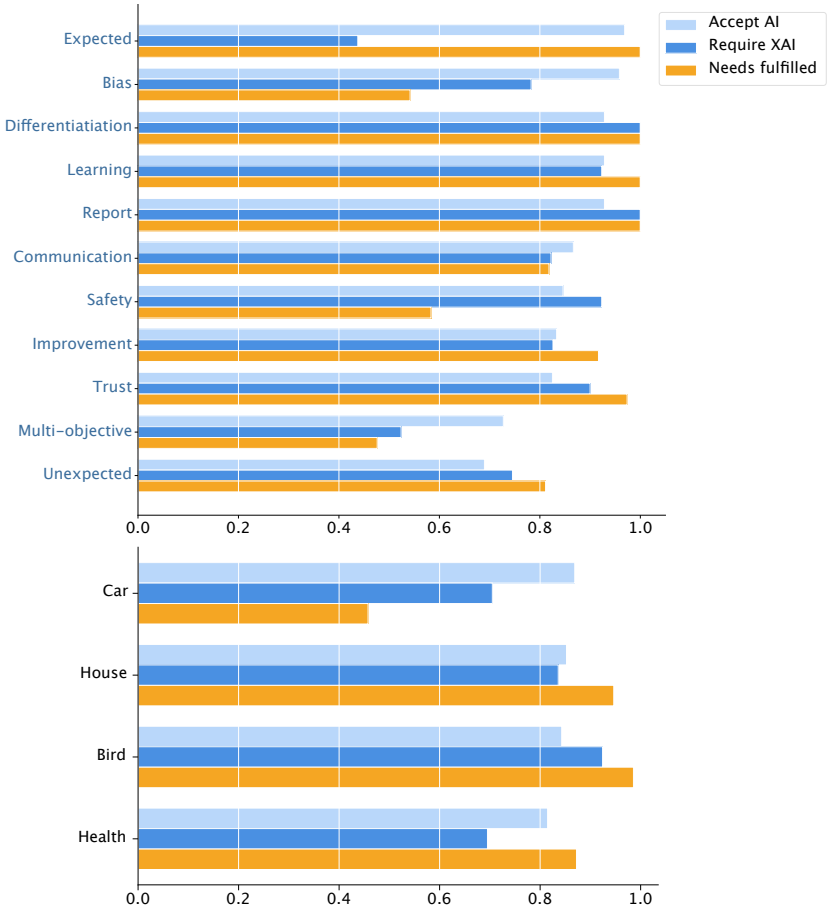


Fig. 4. The rates of **Accept AI**, **Require XAI**, and **Needs fulfilled by the selected explanatory forms** for the explanation goals (top) and tasks (bottom).

2.2.1 Accept AI as decision-support . 84% of the total responses (252/300) were willing to accept AI as decision support. Participants considered the main benefits of using AI were its expertise, convenience, reduce human error, *"boost (user's decision-making) confidence"* (P03), and combine the strengths of human (good at communicating and negotiating with other people) and AI (good

at accurate predictions) (P30). Yet they *"don't want to rely on it 100%"* (P30), and *"look at them (AI) as supplements rather than replacements (of the human)"* (P13). In cases where AI was not acceptable to be involved in decision making, the main concerns were the privacy and safety issues, and because it is a new and unfamiliar technology. The fine-grained rates of "Accept AI" and 'Require XAI" among the 11 explanation goals and 4 tasks are shown in Fig. 4.

We performed Chi-square tests to examine whether participants' demographics or attitude towards AI is related to their willingness to "Accept AI". Among the age, gender, and educational level, only *educational level* was associated with their willingness to "Accept AI", $\chi^2(2, N = 300) = 8.647$, $p = .013$. Participants who had a higher educational level were more acceptable to AI for critical decision-support, and the acceptance rates for each educational level are: *secondary education*: 75%, *undergraduate*: 85%, *postgraduate*: 92%. Their willingness to "Accept AI" in specific tasks and explanation goals did not differ by their familiarity with AI or general attitude toward AI.

2.2.2 Require XAI. 78% of the total responses (228/293) required additional information/explanations from AI. Participants who did not require AI's explanation held two extreme attitudes: some accepted the "black-box" AI thus only required a prediction, and some discredited AI and did not want to take its prediction or any further explanations.

The Chi-square tests showed that among the demographic factors of age, gender, and educational level, only *age* has a significant relationship with "Require XAI", $\chi^2(2, N = 293) = 13.239$, $p = .001$. Participants who were above 55 years old were less likely to "Require XAI" (57%) than those younger than 55 (82%). The likelihood of "Require XAI" did not differ by participants' familiarity with AI. But we observed a significant relationship between their *attitude* toward AI and "Require XAI", $\chi^2(2, N = 283) = 23.739$, $p = .00003$. Participants who had a *neutral* attitude toward AI were the least likely (48%) to check AI's explanations than others (*positive*: 79%, *negative*: 72%, *mixed*: 88%).

2.2.3 Combination of explanatory forms to fulfill XAI needs. After the card sorting, we collected participants' responses to the question "do you think the combination of selected explanatory forms can fulfill your need for the current task and explanation goal?" We ignored the responses who were willing to accept "black-box" AI and did not bother to check explanations (9 out of 288). 83% responses (231/279) would rate their needs had been fulfilled by the prototyping cards combination. The fine-grained rates of "Needs fulfilled" among the 11 explanation goals and 4 tasks are shown in Figure 4.

We performed Chi-square tests to identify if the rate of "Needs fulfilled" differs by demographic factors or attitudes toward AI. *Age* had a significant relationship with the "Needs fulfilled" rate, $\chi^2(2, N = 279) = 13.637$, $p = .001$. Participants who were above 55 years old were less likely to rate "Needs fulfilled" (63%) than those younger than 55 (85%). *Educational level* also had a significant relationship with "Needs fulfilled" rate, $\chi^2(2, N = 279) = 9.534$, $p = .008$, and the rates for educational levels were: *secondary education*: 92%, *undergraduate*: 76%, *postgraduate*: 83%. The rate of "Needs fulfilled" did not differ by participants' gender, or familiarity with AI. But we observed a significant relationship between their *attitude* toward AI and "Needs fulfilled", $\chi^2(2, N = 270) = 26.985$, $p = 1.0^{-5}$. Participants who had a *negative* attitude toward AI would less likely (40%) rate their needs had been fulfilled than others (*mixed*: 92%, *neutral*: 76%, *positive*: 82%).

2.3 Card Selection Results: Preferred Explanatory Forms for Each Explanation Goal

In the next two sections, we will present the quantitative analysis results on card selection and card sorting respectively. A total of 248 valid card selection & sorting data were collected. For card selection, we analyzed participants' preferences of explanatory forms for each explanation goal. The aggregated results are shown in Fig. 5. 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 explanatory form rule.

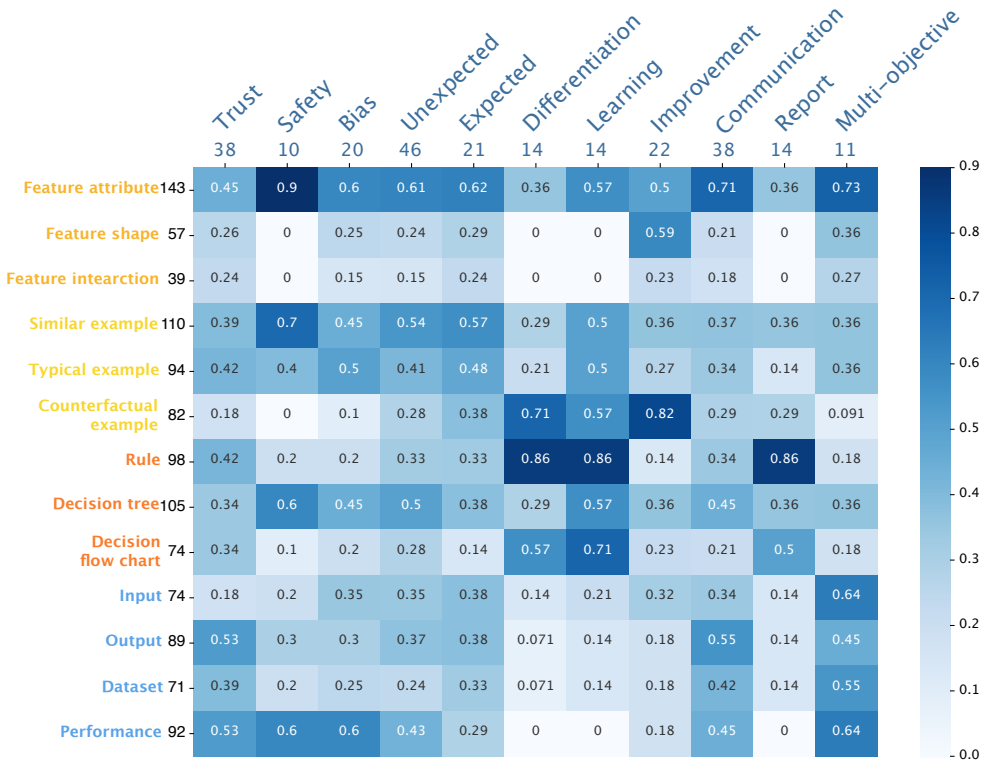
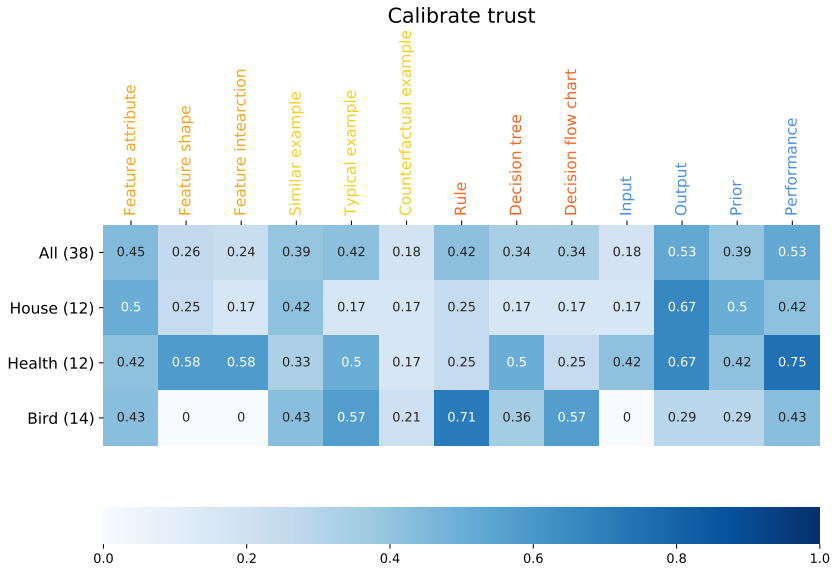
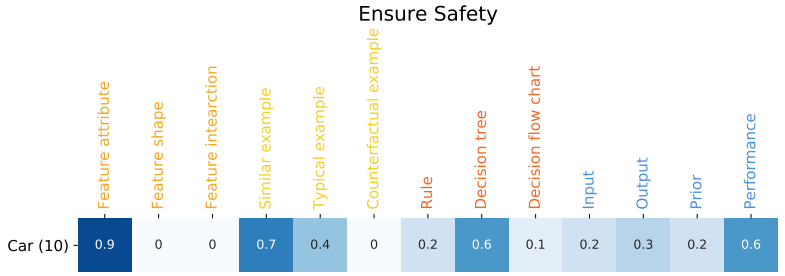


Fig. 5. **The explanatory form—explanation goal matrix heatmap**. The darkness level and number in the grid is the percentage of a explanatory form selected for that explanation goal. The number under each goal (on the horizontal top) is the total number of card selection & sorting data collected for that explanation goal. The number beside each explanatory form (on the vertical left) is the total number of times an explanatory form was selected in the card selection & sorting data.

Calibrating trust. For quantitative results on the most frequently selected explanatory 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.





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)

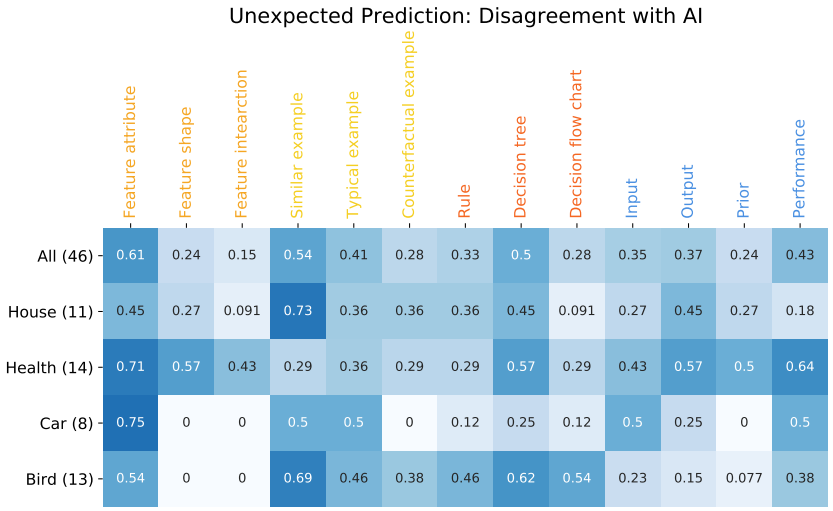
Similar example (7/10) were preferred since it showed *"what's the condition or what kind of decision the car gonna make"* (P32), although participants did not focus on its similarity nature, but rather assumed it can showcase a variety of cases including the extreme cases. Several participants chose 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 [2] 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)

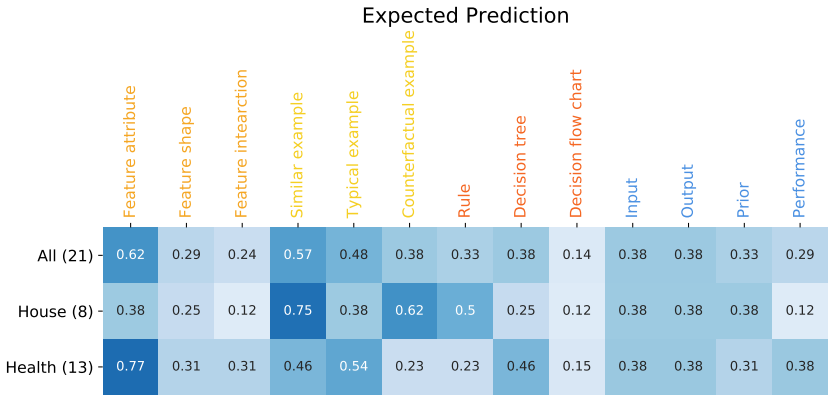
Feature attribution (12/20) was also chosen since participants wanted to check if AI could still detect important features in minority conditions.

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

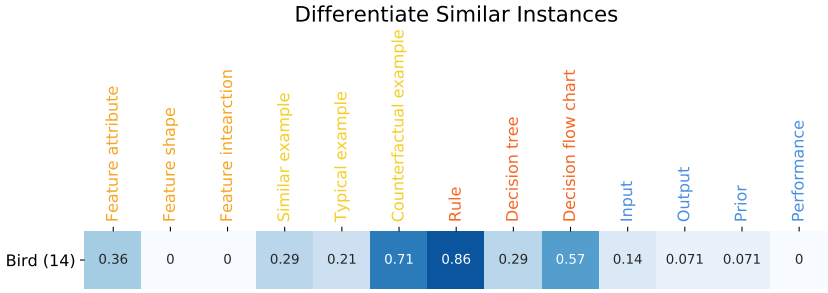


Unexpected Prediction: When Users Disagree with AI. The frequently selected explanatory forms are: feature attribution (28/46), similar example (25/46), decision tree (23/46), and performance (20/46).

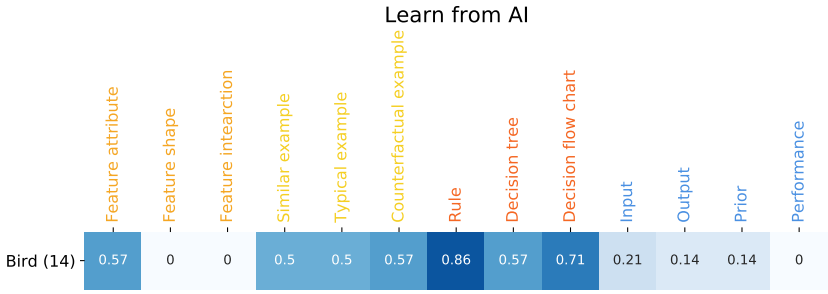
Despite users disagree with AI, if users' judgment is included in AI's differential prediction list or range, users would think AI has the ability to discern similar predictions, and may resolve the prediction disagreement to an extent, as some participants suggested: *"What would be really interesting 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 explanatory forms for this explanation goal unexpected.



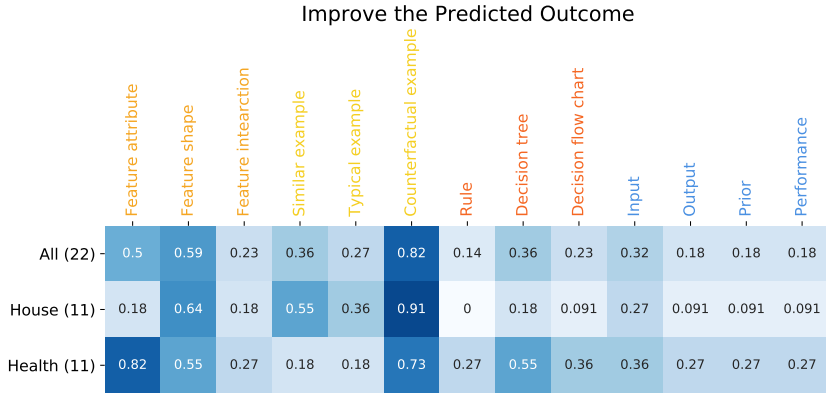
Expected Prediction: When Users Agree with AI. The frequently selected explanatory forms are: feature attribution (13/21), similar example (12/21), and typical example (10/21).



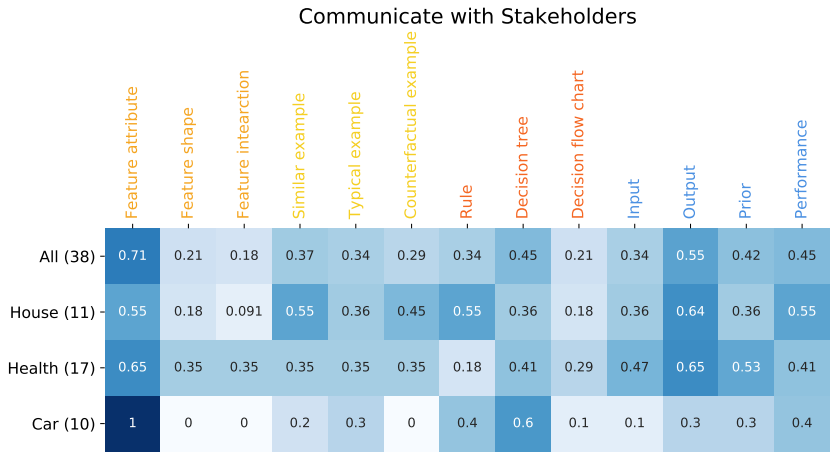
Differentiating similar instances. Rule (12/14) and counterfactual example (10/14) were the most preferable forms. Participants chose rule since *"you could write that you differentiated the bird's tail were long or short, or beak thin or thick"* (P10). The counterfactual examples *"identify where specifically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



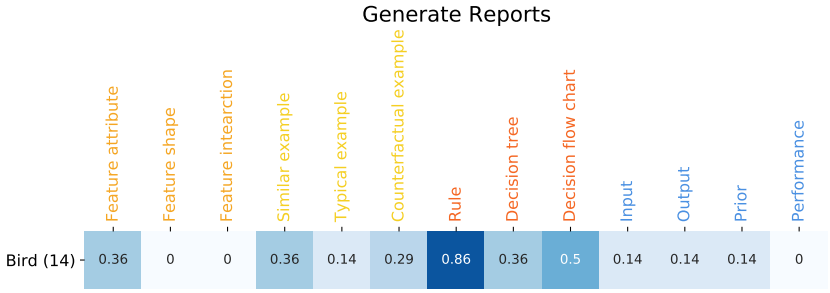
Learning from AI. Rule-based explanations (rule: 12/14, decision flow chart: 10/14, decision tree: 8/14) were more favourable for the explanation goal to learn, since they showed *"a learning process. It has like how you could recognize a bird. So help me to learn some new knowledge"* (P02). *"(decision 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 [1]).



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 explanatory forms.

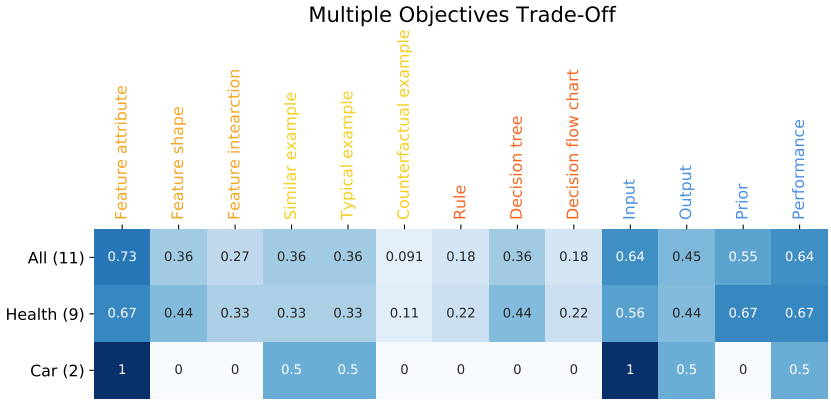
Rule were selected because its text description format can conveniently generate text reports. *"I have to write the explanation"* (P08, P09); *"You can not only by looking at the images and get some explanation. You need some more specific description."* (P08)

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 explanatory 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 explanatory types were 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).

Fig. 6. **Clusters of the explanation goals** The explanation goals that are close to each other indicate they have similar patterns on participants' explanatory 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 explanatory 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.

To gain an intuitive understanding of how similar the explanatory forms are to each other, we applied multidimensional scaling and visualized their similarities as distances on a 2D scatter plot as well as a dendrogram and pairwise similarity heatmap in Figure 7. The similarity was measured as the co-occurrence of a pair of explanatory forms in card selection. Based on the 2D positions of the explanatory forms, we applied k-means and hierarchical clustering analysis and yielded similar clustering patterns.

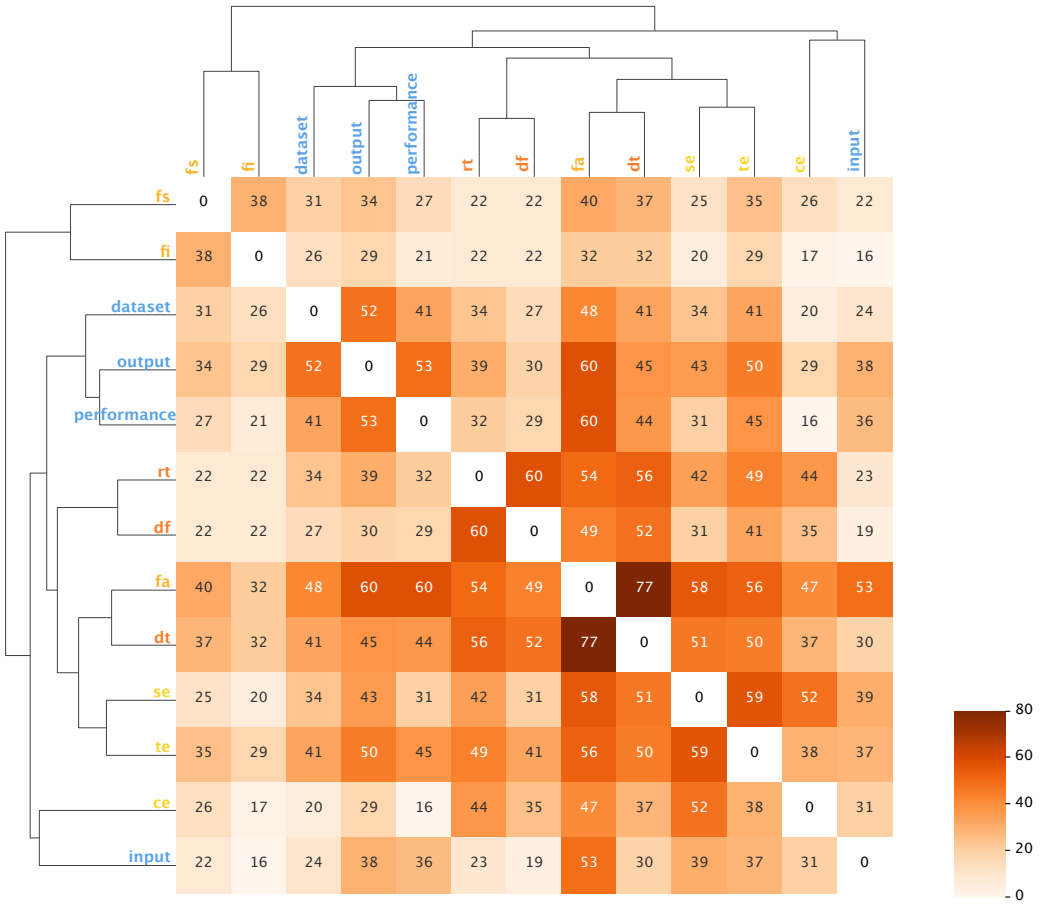


Fig. 7. **Similarity matrix and dendrogram of the 12 end-user-friendly explanatory forms**. The pairwise similarities are measured by co-occurrence of two cards selected in a card selection & sorting response. Darker orange indicates the two pairs are more likely to be selected in the same XAI card combination, and the co-occurrence numbers are shown in each grid. The dendrogram was generated using hierarchical clustering.

2.4 Card Sorting Results on Explanatory 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 explanatory forms, the rest of the explanation goals had a consistent pattern ($p < 0.05$). Table 1 summarizes the aggregated sorting for explanation goals that have significant consistent patterns.

Table 1. **Aggregated sorting of explanatory 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 explanatory 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 explanatory 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 explanatory 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 2.

Table 2. **Aggregated sorting of explanatory 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 explanatory 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 explanatory 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 explanatory 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 explanatory card selection; whereas card selection preferences do differ by age group, educational level, familiarity with AI, and attitude towards AI ($p < 0.05$).

3 PARTICIPANTS' INFORMATION

Participant number	Age	Sex	Educational 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

Participant number	Age	Sex	Educational level	Major or Industry	AI familiarity	AI attitudes	Tasks	Interview duration (min)
P14	47	M	Some college credit	healthcare-sterilization work	never	interested	House; Bird	55
P15	73	M	Professional	retired	heard	skeptical	Car	81
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

Participant number	Age	Sex	Educational level	Major or Industry	AI familiarity	AI attitudes	Tasks	Interview duration (min)
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
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

4 STUDY MATERIALS

4.1 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

4.2 Interview materials

The next few pages contain the following interview materials, used in the study:

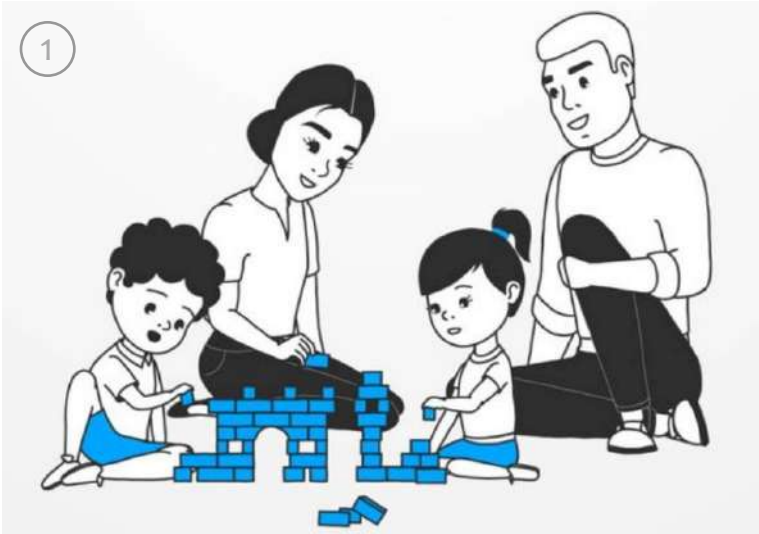
(1) The four tasks shown as storyboards;
(2) The explanation goals shown as storyboards;
(3) The explanatory forms generated by the EUCA framework, shown as cards.

REFERENCES

[1] Robert R. Hoffman and Gary Klein. 2017. Explaining Explanation, Part 1: Theoretical Foundations. *IEEE Intelligent Systems* 32, 3 (2017), 68–73. https://doi.org/10.1109/MIS.2017.54

[2] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A Survey on Bias and Fairness in Machine Learning. (2019). arXiv:1908.09635 http://arxiv.org/abs/1908.09635

Selling your house

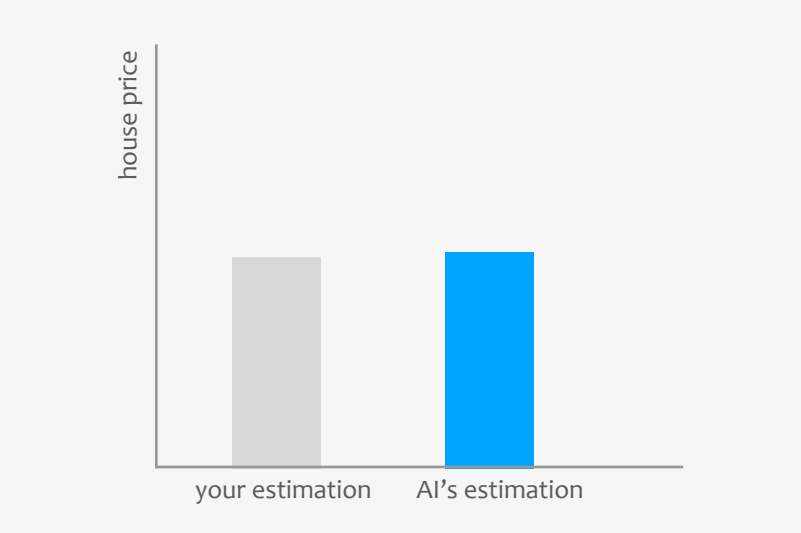


Suppose your family is expanding

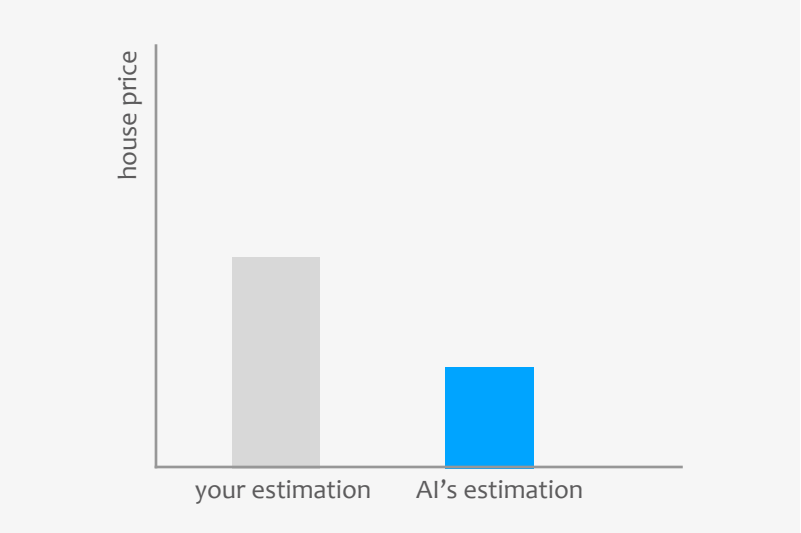
and you need to sell your current house, for a bigger one.

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

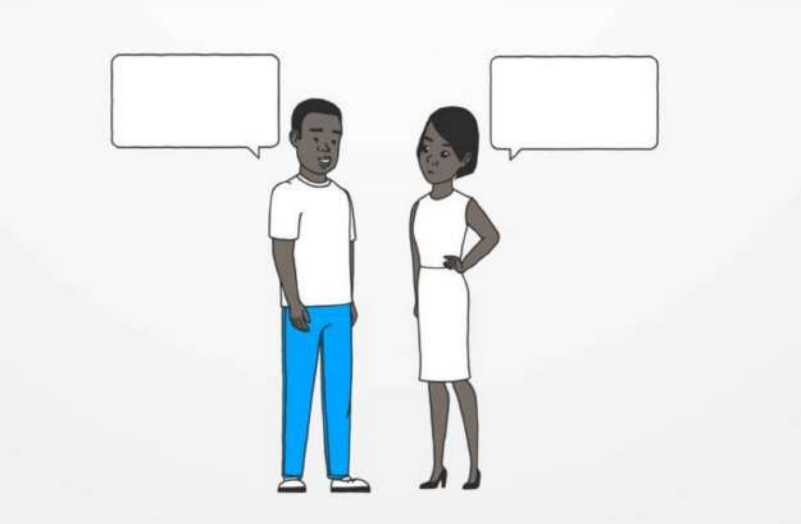
You get to know there is an **aritifical intelligence (AI)** tool that can **predict house price**. It may help you to get a propoer estimate of your house.



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



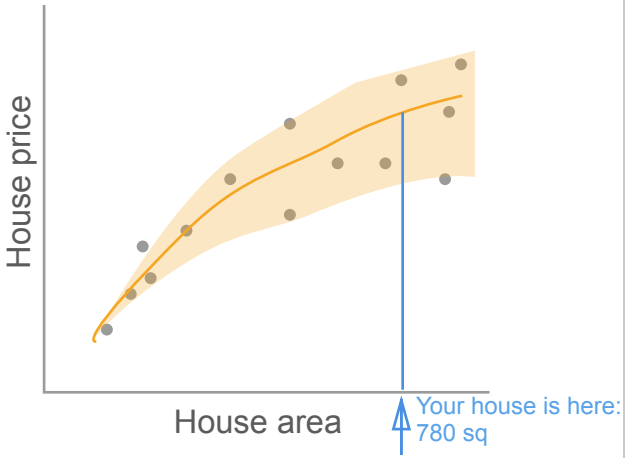
AI's prediction does ***not align*** with your estimation

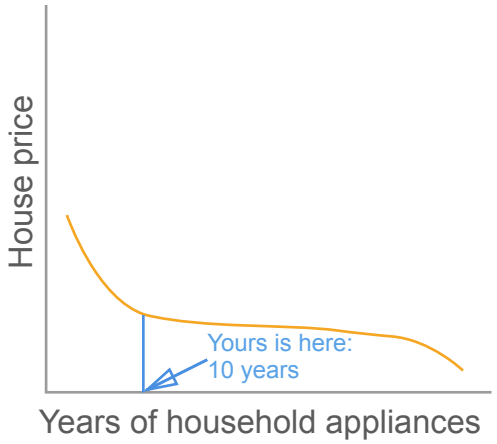


You need to ***communicate*** your decision with your family

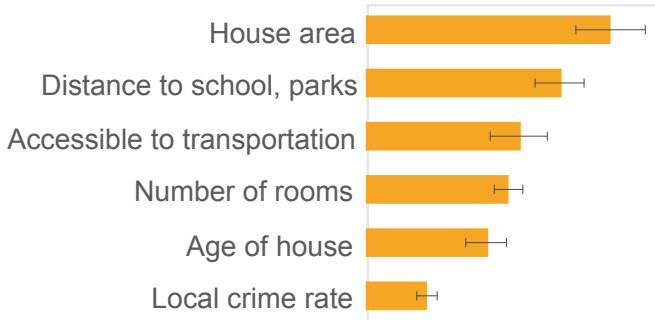
You doubt whether to ***trust*** the AI tool or not

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.

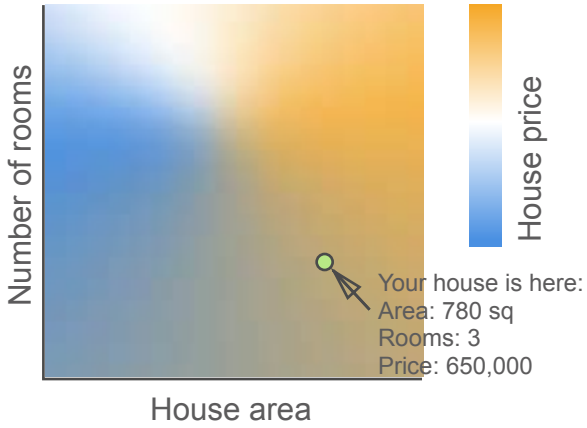




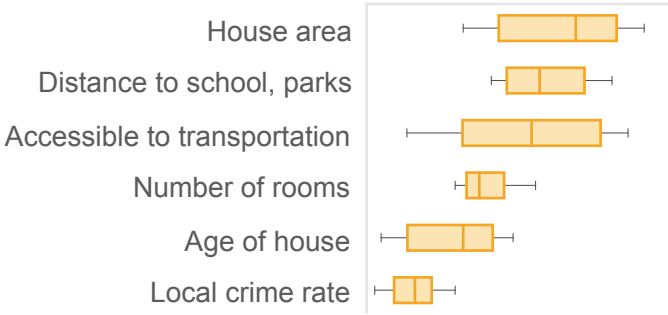
How important is each feature to the result:



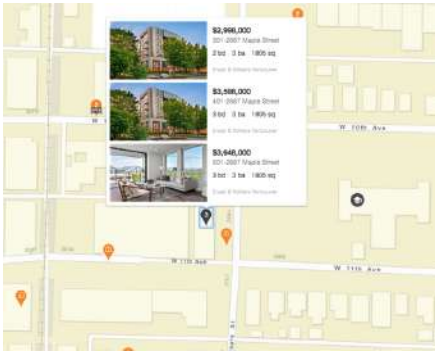
The houses of **similar *price*** as yours



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
- ……

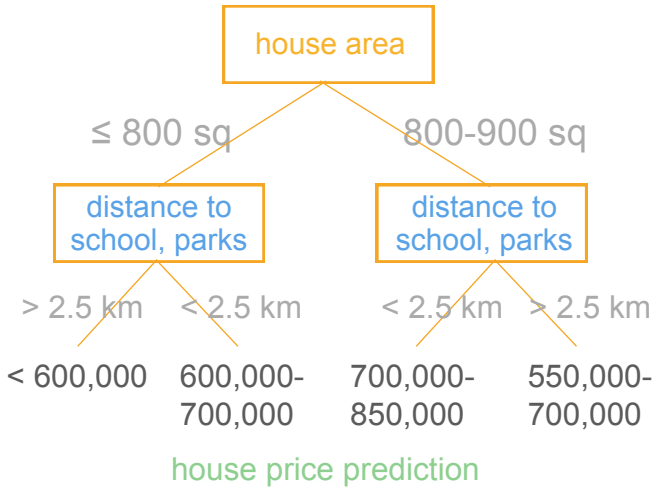
If the feature of your house had changed to the following feature,
your house price would have increased by 10%:

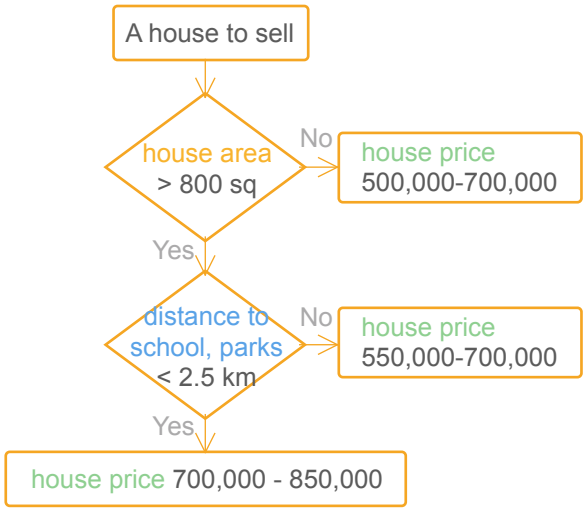
- have a back yard, or
- 3 bathrooms, or
- 1200 sq, or
- less than 10 years old, or
- has new household appliances
- ……

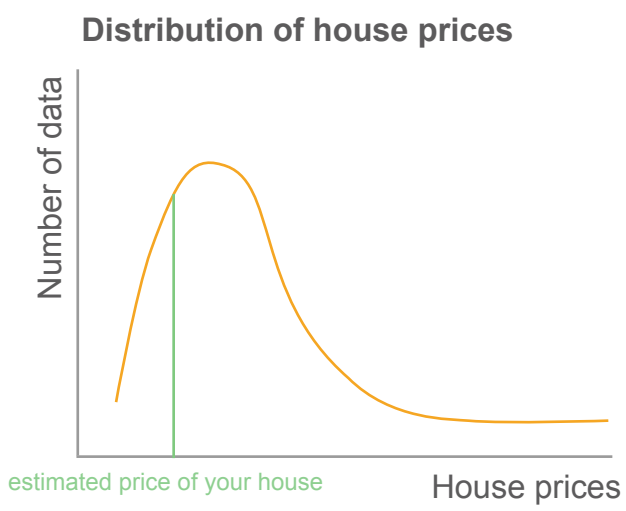
If **house area** ≤ 800 sq,
and **distance to school, parks** > 2.5 km,
Then house price **is no more than 600,000**

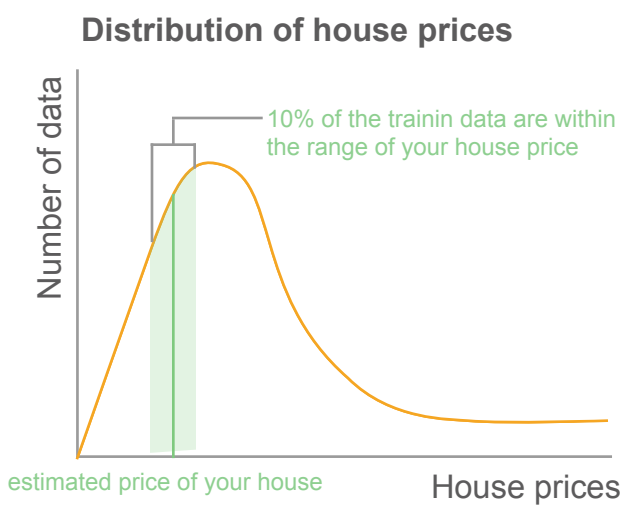
If **house area** is 800 - 900 sq,
and **distance to school, parks** < 2.5 km,
Then house price **is about 700,000-850,000**

	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

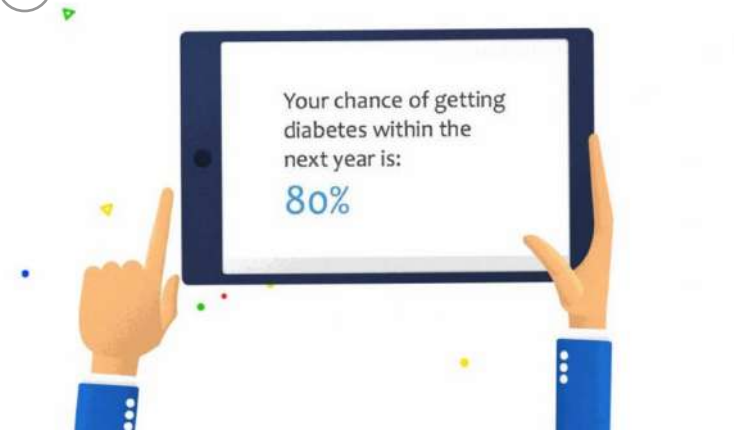
1

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.

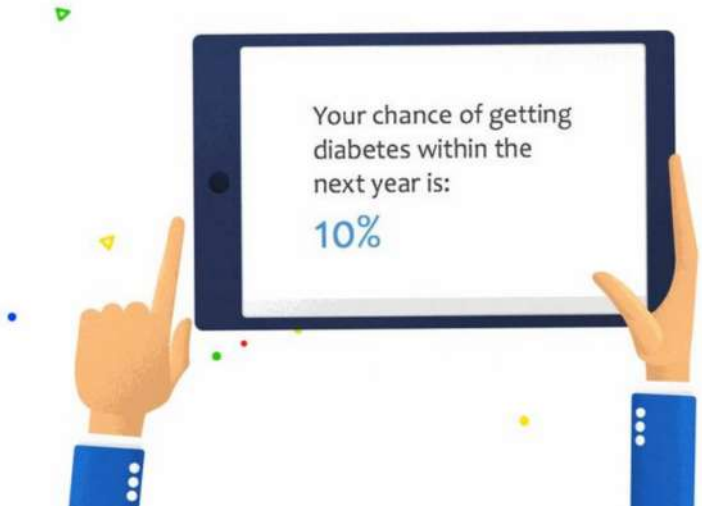
4

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***.



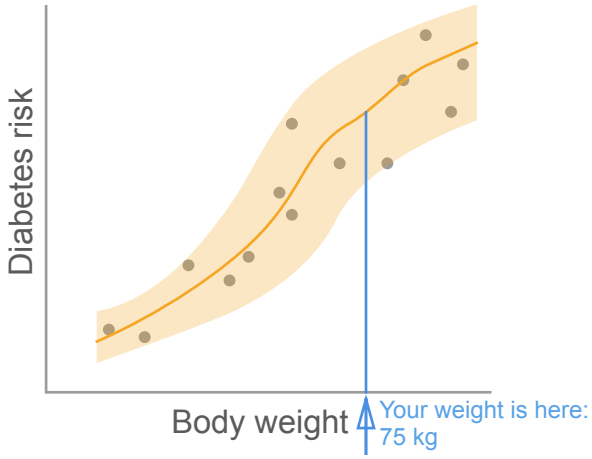
It predicts your chance of getting diabetes is ***low***.

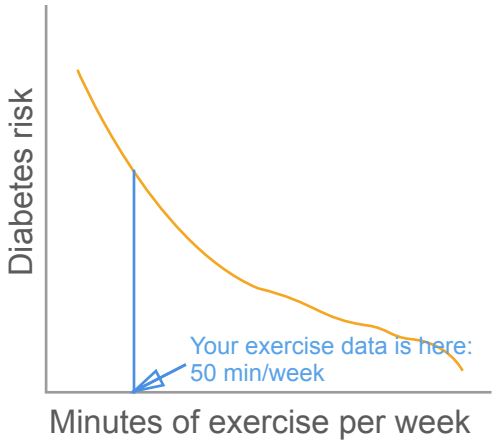
Diabetes tends to run in your family, and you're afraid of getting it someday.

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

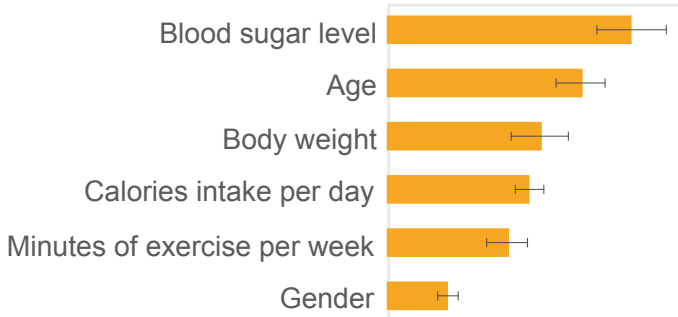
You're aware that the insurance company may use such a prediction from the software to ***determine your insurance fee and benefits***.

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



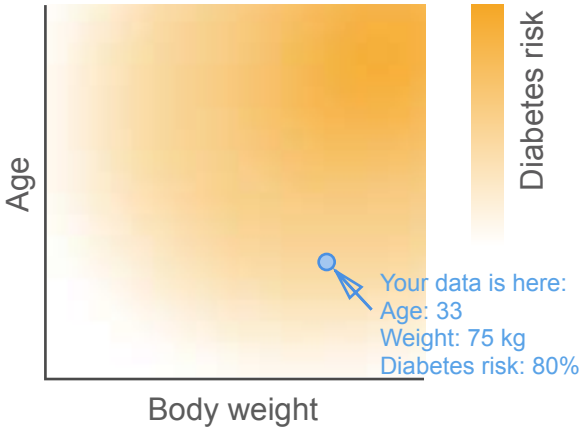


How important is each feature to the result:

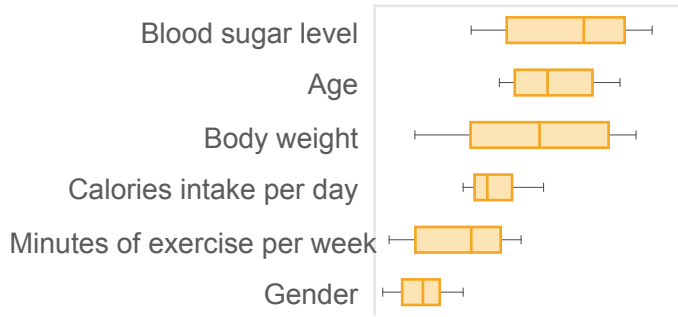


The case that has the **similar *diabetes risk*** as yours:

- 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: ……
- .....

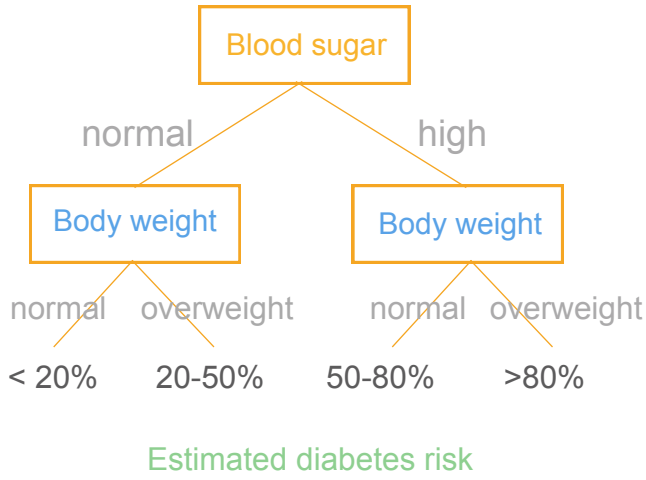
If your health data had changed to the following, your diabetes risk would have decreased by 20%:

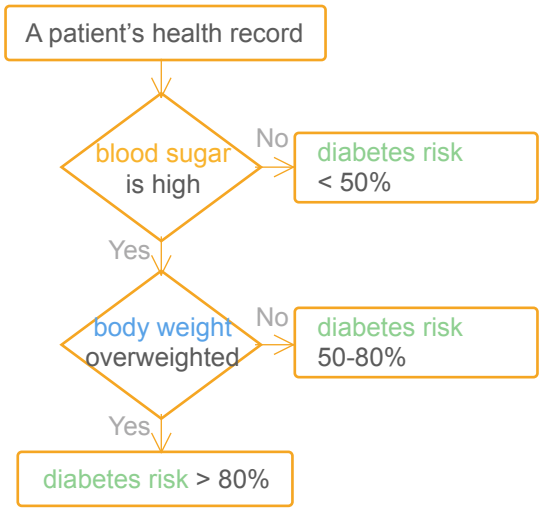
- 3 years younger than now
- Body weight: loss 5 kg
- Increase 50 min of weekly exercise
- Reduce 500 calories of daily calories intake
- ……

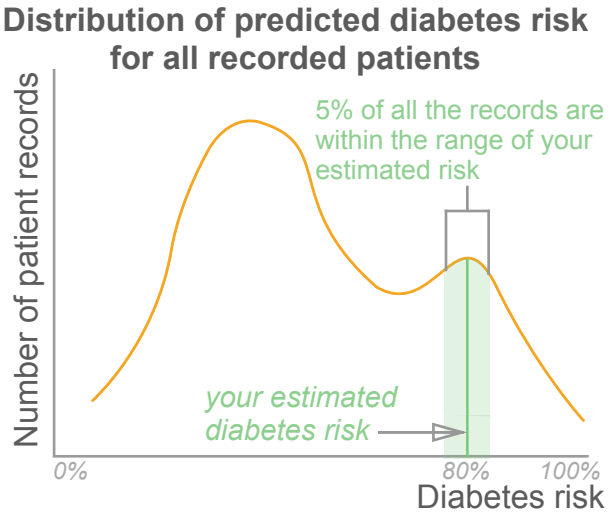
If **blood sugar** is **high**,
and **body weight** is **overweighted**,
Then the estimated diabetes risk
is above 80%

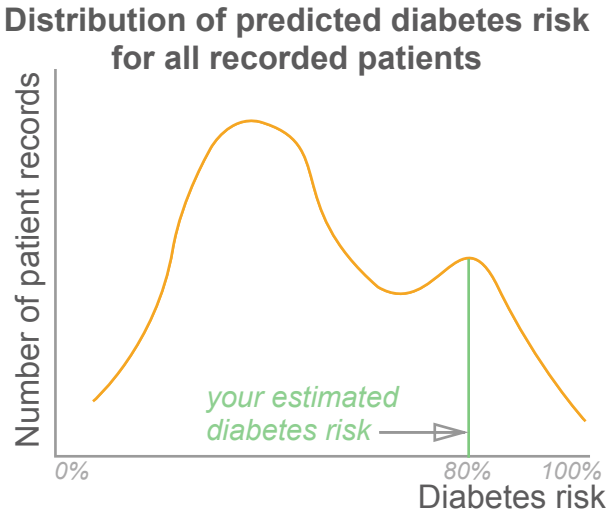
If **blood sugar** is **normal**,
and **body weight** is **overweighted**,
Then the estimated diabetes risk
is about 20-50%

	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.***

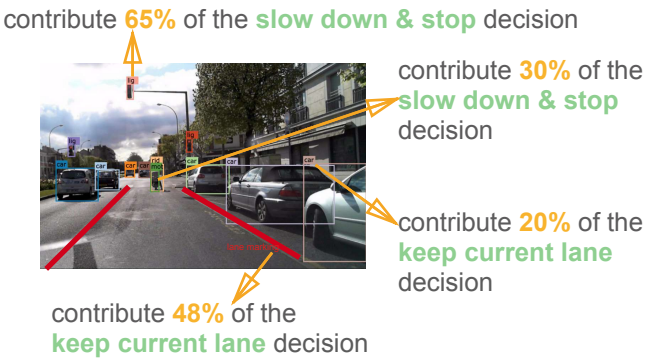
You need to know whether the autopilot mode is ***safe*** and ***reliable.***

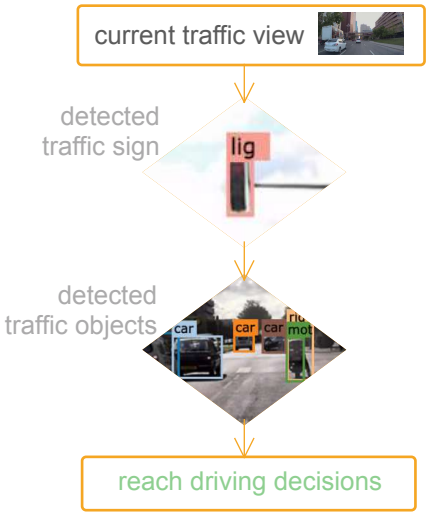
Similar traffic conditions as the current one, from the dataset to train the self-driving car:

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:

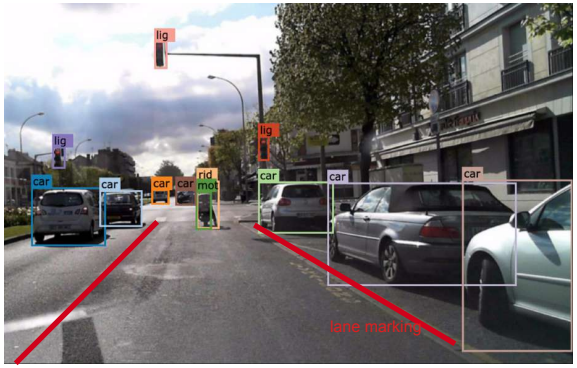




The current driving decisions, and **their percentage in the training dataset** where the self-driving car learns from

	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:

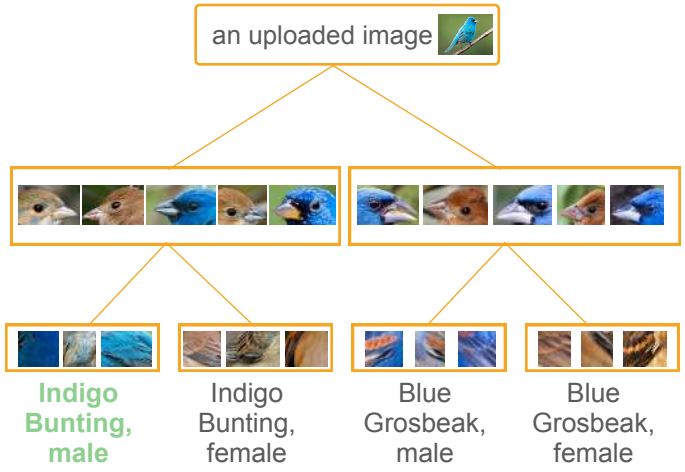
*Measured using average distance driven between disengagements**

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

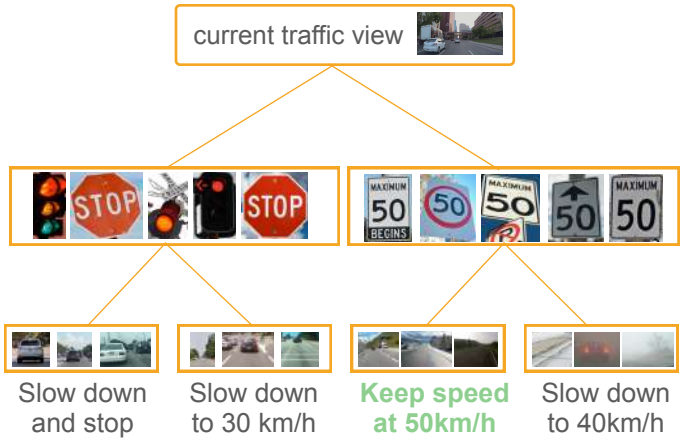
* Disengagement means when the automated system is switched off by the intervention of a human driver

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

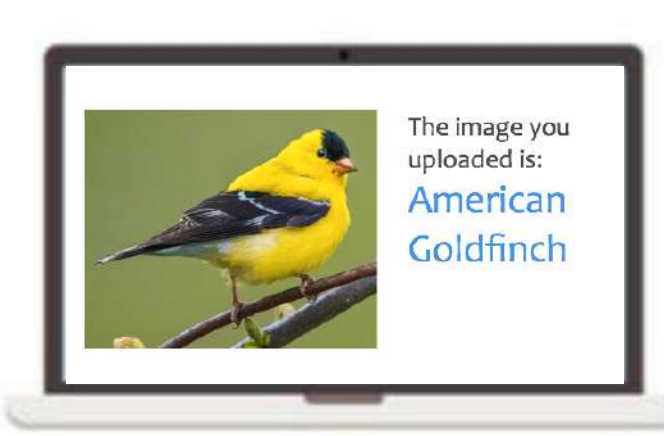
If **traffic sign** is **50km/h speed limit**,
and the speed of the **car in front** are
the **same or faster**,
Then the speed is kept at
50km/h

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%

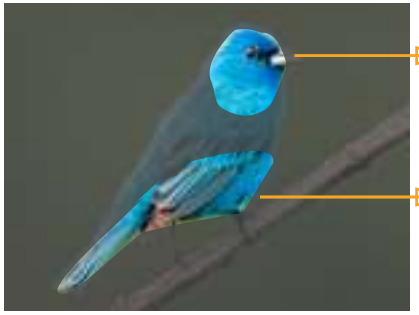
The three most likely bird according to your uploaded image, and **typical examples**

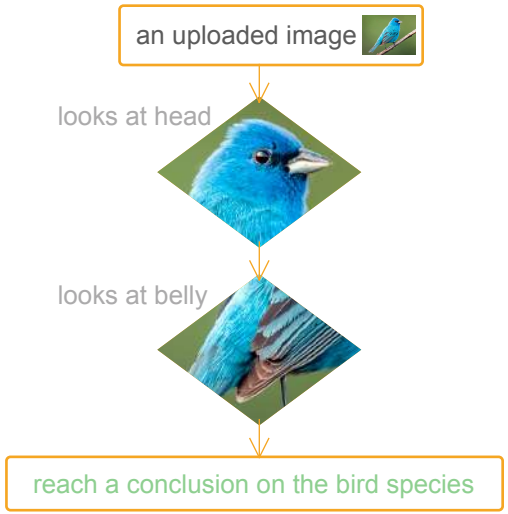
Indigo Bunting
95%

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%