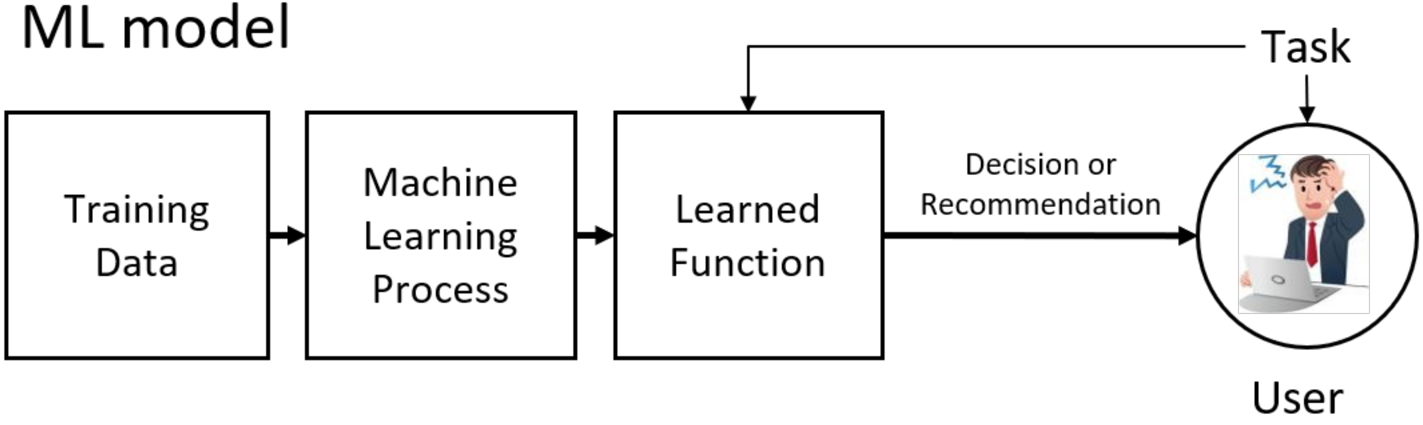
Exploring the Future of Explainable AI (XAI) in HCI

What is XAI?

"Explainable AI (XAI) is a set of processes and methods that allows human users to comprehend and trust the results and outputs of machine learning algorithms."

"Should I marry Ann?"
_David
"NO!"
_Oracle

ML model



Training Data

Machine Learning Process

Learned Function

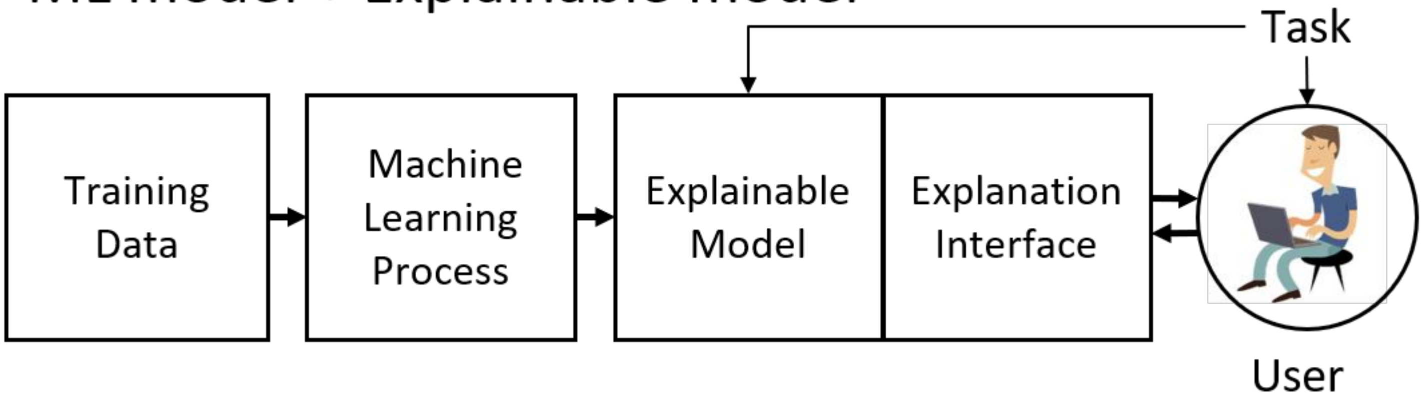
Decision or Recommendation

Task

User

- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

ML model + Explainable model



Training Data

Machine Learning Process

Explainable Model

Explanation Interface

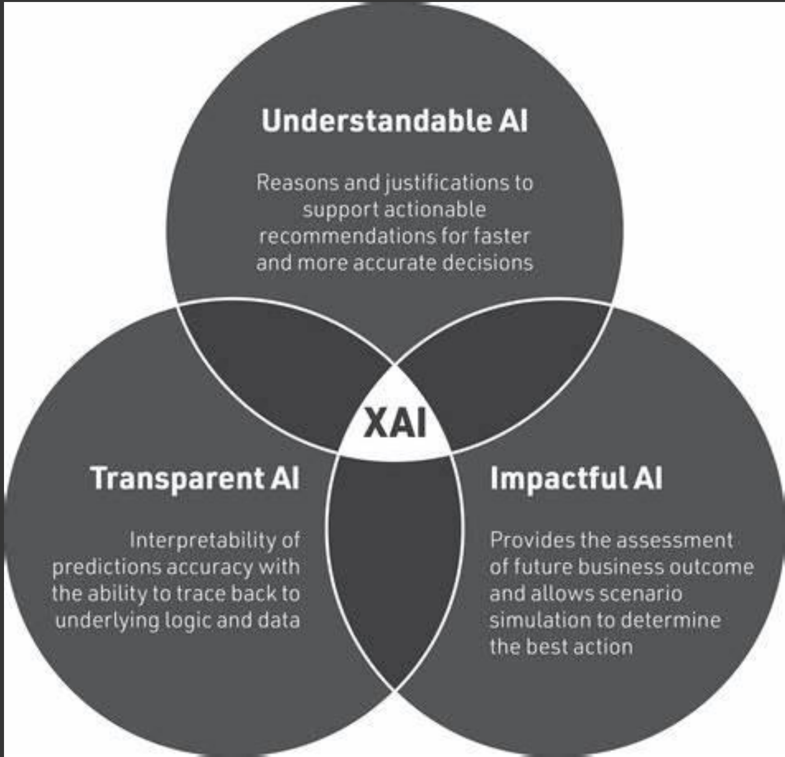
Task

User

- I understand why
- I understand why not
- I know when you succeed
- I know when you fail
- I know when to trust you
- I know why you erred

The Emergence of XAI

- "Transition from opaque, black-box AI models to systems that are interpretable and understandable."
- "Driven by the need for trust in AI in critical areas like healthcare, finance, and law enforcement."



Understandable AI

Reasons and justifications to support actionable recommendations for faster and more accurate decisions

XAI

Transparent AI

Interpretability of predictions accuracy with the ability to trace back to underlying logic and data

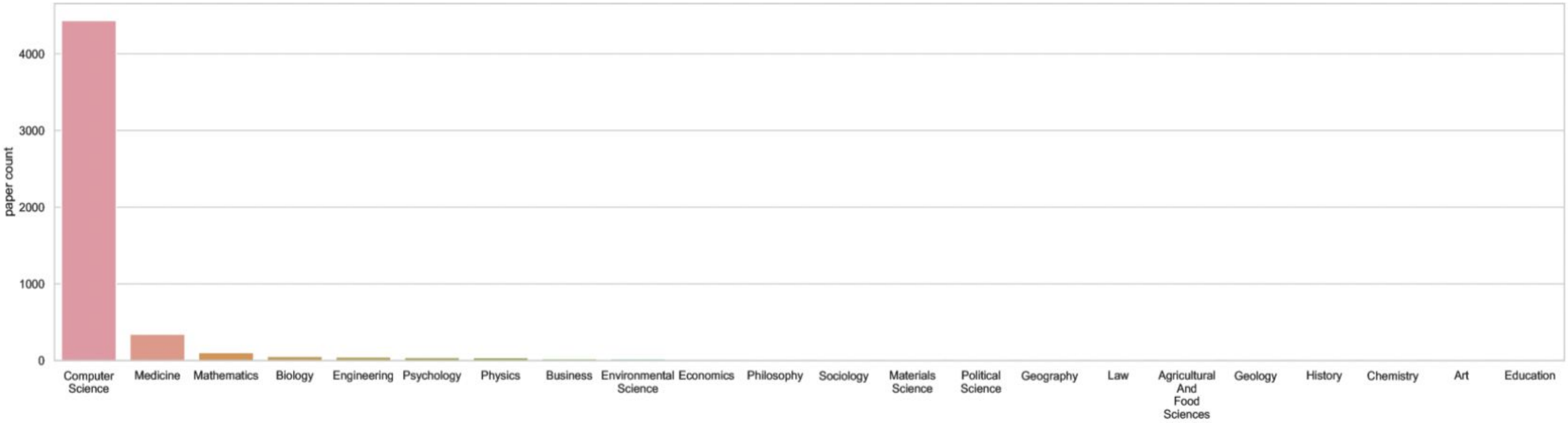
Impactful AI

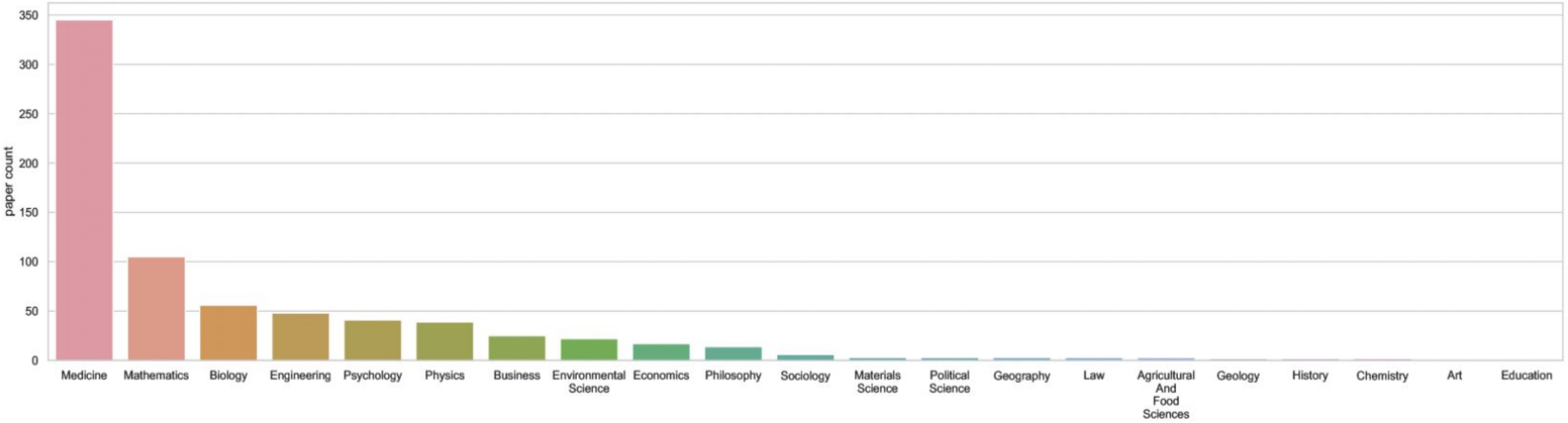
Provides the assessment of future business outcome and allows scenario simulation to determine the best action

XAI in Research - Recent Trends

- Expansion Beyond Computer Science:
  XAI research has grown notably in non-CS fields, especially in 2016, 2018, and 2021.
  Significant advancements in applying XAI principles in medicine, psychology, and engineering.

- Cross-Field Influence:
  Diverse citing relationships between XAI and other fields.
  For example, XAI in computer science cites psychology more often than vice versa, but this pattern changes with fields like medicine.

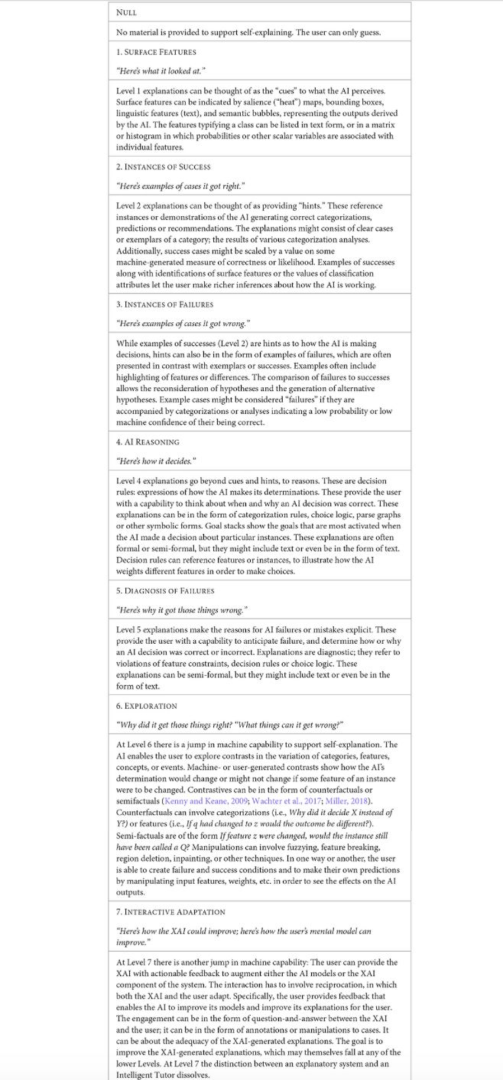

paper count
4000
3000
2000
1000
0
Computer Science
Medicine
Mathematics
Biology
Engineering
Psychology
Physics
Business
Environmental Science
Economics
Philosophy
Sociology
Materials Science
Political Science
Geography
Law
Agricultural And Food Sciences
Geology
History
Chemistry
Art
Education


paper count
350
300
250
200
150
100
50
0
Medicine
Mathematics
Biology
Engineering
Psychology
Physics
Business
Environmental Science
Economics
Philosophy
Sociology
Materials Science
Political Science
Geography
Law
Agricultural And Food Sciences
Geology
History
Chemistry
Art
Education

XAI Measurement Science - The "Scorecard" Method

Front. Comput. Sci., 09 May 2023, Robert Hoffman et. al

- The Scorecard provides a judgment scale to assess how well explanations support user sensemaking, from basic information to in-depth cognitive support
- Emphasizes the importance of **user-centric explanations in AI,** advocating for designs that are understandable and cognitively valuable to diverse users



NULL
No material is provided to support self-explaining. The user can only guess.
1. SURFACE FEATURES *"Here's what it looked at."*
Level 1 explanations can be thought of as the "cues" to what the AI perceives. Surface features can be indicated by salience ("heat") maps, bounding boxes, linguistic features (text), and semantic bubbles, representing the outputs derived by the AI. The features typifying a class can be listed in text form, or in a matrix or histogram in which probabilities or other scalar variables are associated with individual features.
2. INSTANCES OF SUCCESS *"Here's examples of cases it got right."*
Level 2 explanations can be thought of as providing "hints." These reference instances or demonstrations of the AI generating correct categorizations, predictions or recommendations. The explanations might consist of clear cases or exemplars of a category; the results of various categorization analyses. Additionally, success cases might be scaled by a value on some machine-generated measure of correctness or likelihood. Examples of successes along with identifications of surface features or the values of classification attributes let the user make richer inferences about how the AI is working.
3. INSTANCES OF FAILURES *"Here's examples of cases it got wrong."*
While examples of successes (Level 2) are hints as to how the AI is making decisions, hints can also be in the form of examples of failures, which are often presented in contrast with exemplars or successes. Examples often include highlighting of features or differences. The comparison of failures to successes allows the reconsideration of hypotheses and the generation of alternative hypotheses. Example cases might be considered "failures" if they are accompanied by categorizations or analyses indicating a low probability or low machine confidence of their being correct.
4. AI REASONING *"Here's how it decides."*
Level 4 explanations go beyond cues and hints, to reasons. These are decision rules: expressions of how the AI makes its determinations. These provide the user with a capability to think about when and why an AI decision was correct. These explanations can be in the form of categorization rules, choice logic, parse graphs or other symbolic forms. Goal stacks show the goals that are most activated when the AI made a decision about particular instances. These explanations are often formal or semi-formal, but they might include text or even be in the form of text. Decision rules can reference features or instances, to illustrate how the AI weights different features in order to make choices.
5. DIAGNOSIS OF FAILURES *"Here's why it got those things wrong."*
Level 5 explanations make the reasons for AI failures or mistakes explicit. These provide the user with a capability to anticipate failure, and determine how or why an AI decision was correct or incorrect. Explanations are diagnostic; they refer to violations of feature constraints, decision rules or choice logic. These explanations can be semi-formal, but they might include text or even be in the form of text.
6. EXPLORATION *"Why did it get those things right? "What things can it get wrong?"*
At Level 6 there is a jump in machine capability to support self-explanation. The AI enables the user to explore contrasts in the variation of categories, features, concepts, or events. Machine- or user-generated contrasts show how the AI's determination would change or might not change if some feature of an instance were to be changed. Contrastives can be in the form of counterfactuals or semifactuals (Kenny and Keane, 2009; Wachter et al., 2017; Miller, 2018). Counterfactuals can involve categorizations (i.e., *Why did it decide X instead of Y?)* or features (i.e., *If q had changed to z would the outcome be different?*). Semi-factuals are of the form *If feature z were changed, would the instance still have been called a Q?* Manipulations can involve fuzzying, feature breaking, region deletion, inpainting, or other techniques. In one way or another, the user is able to create failure and success conditions and to make their own predictions by manipulating input features, weights, etc. in order to see the effects on the AI outputs.
7. INTERACTIVE ADAPTATION *"Here's how the XAI could improve; here's how the user's mental model can improve."*
At Level 7 there is another jump in machine capability: The user can provide the XAI with actionable feedback to augment either the AI models or the XAI component of the system. The interaction has to involve reciprocation, in which both the XAI and the user adapt. Specifically, the user provides feedback that enables the AI to improve its models and improve its explanations for the user. The engagement can be in the form of question-and-answer between the XAI and the user; it can be in the form of annotations or manipulations to cases. It can be about the adequacy of the XAI-generated explanations. The goal is to improve the XAI-generated explanations, which may themselves fall at any of the lower Levels. At Level 7 the distinction between an explanatory system and an Intelligent Tutor dissolves.

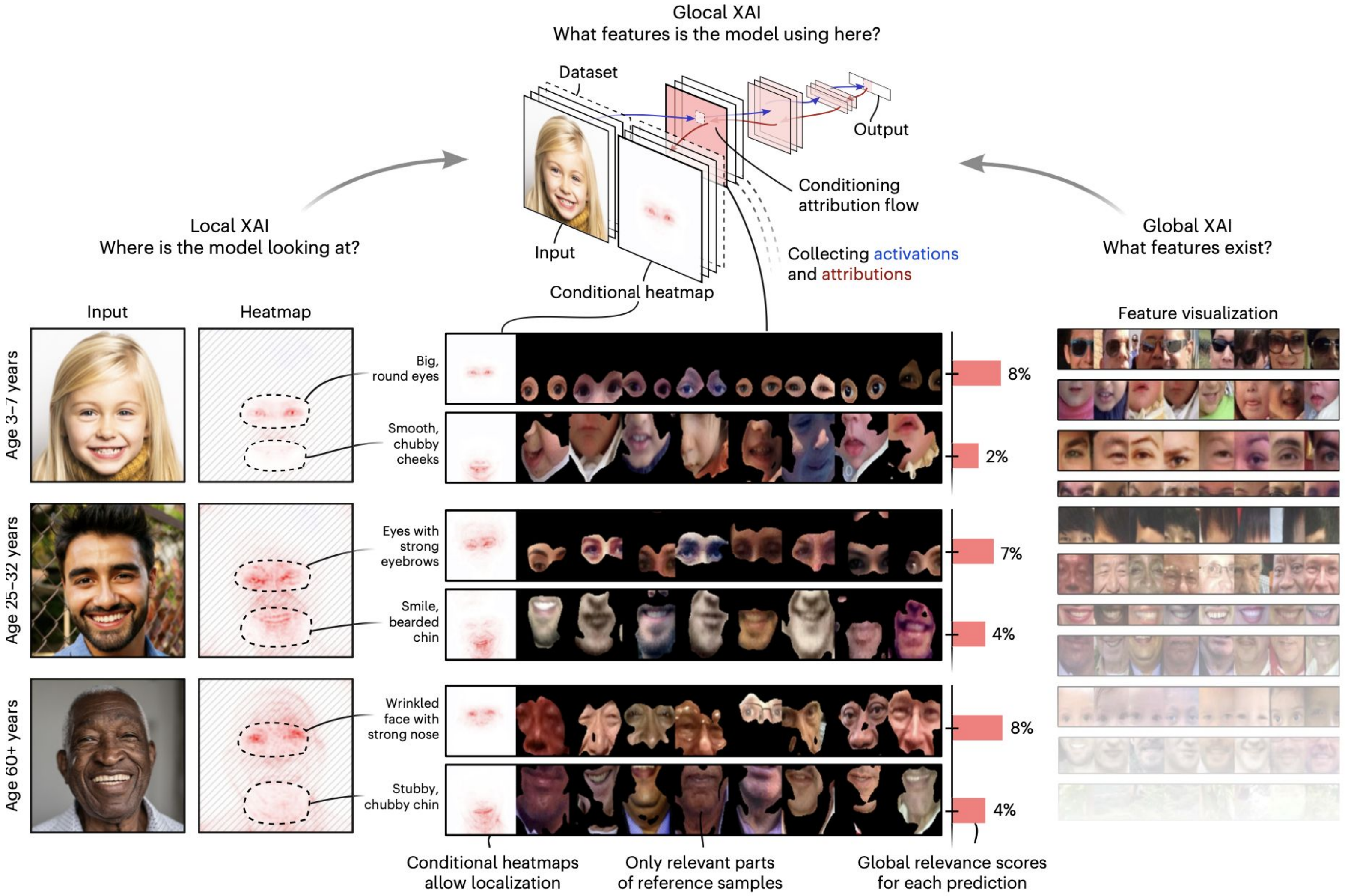
- Guidance for Developers: It provides a framework for developers to extend their machine-generated explanations.

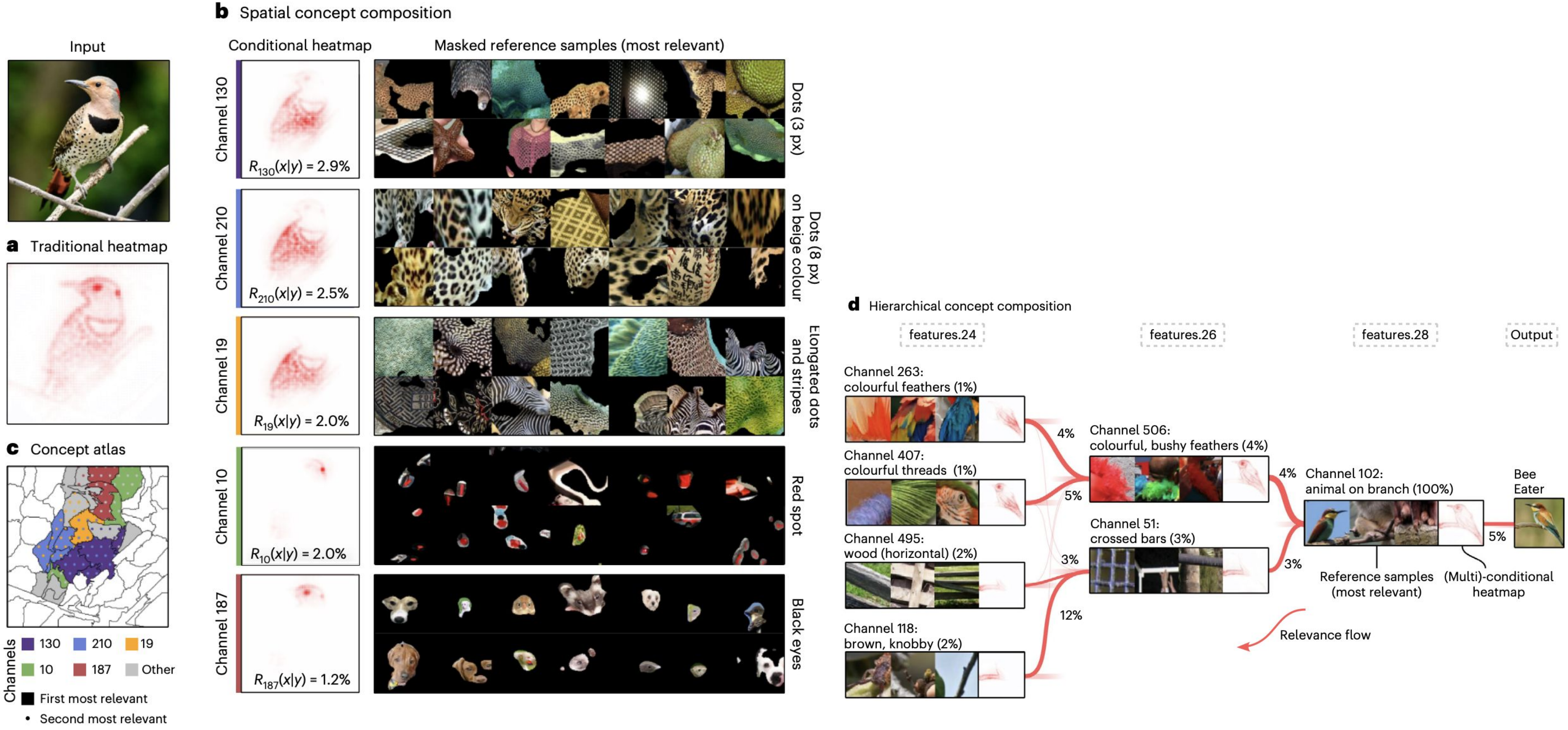
- Future Recommendations: Stressing the need for AI systems to support user sensemaking and cognitive depth in explanations.

Case Study - Concept Relevance Propagation

Developed by Prof. Thomas Wiegand and team, CRP is a method that translates AI decisions into human-understandable concepts.

- involves interpreting AI decisions by identifying not only relevant input features but also the **underlying concepts** and their **specific locations** within the neural network, making the AI's decision-making process more transparent and understandable
- CRP sets a new standard for AI explainability, making it easier for users to understand and interact with AI systems.


Glocal XAI
What features is the model using here?
Dataset
Output
Input
Conditioning attribution flow
Collecting activations and attributions
Conditional heatmap
Local XAI
Where is the model looking at?
Global XAI
What features exist?
Input
Heatmap
Feature visualization
Age 3–7 years
Big, round eyes
8%
Smooth, chubby cheeks
2%
Age 25–32 years
Eyes with strong eyebrows
7%
Smile, bearded chin
4%
Age 60+ years
Wrinkled face with strong nose
8%
Stubby, chubby chin
4%
Conditional heatmaps allow localization
Only relevant parts of reference samples
Global relevance scores for each prediction


Input
a Traditional heatmap
c Concept atlas
Channels
■ 130
■ 210
■ 19
■ 10
■ 187
■ Other
■ First most relevant
• Second most relevant
b Spatial concept composition
Conditional heatmap
Masked reference samples (most relevant)
Channel 130
$R_{130}(x|y)$ = 2.9%
Dots (3 px)
Channel 210
$R_{210}(x|y)$ = 2.5%
Dots (8 px) on beige colour
Channel 19
$R_{19}(x|y)$ = 2.0%
Elongated dots and stripes
Channel 10
$R_{10}(x|y)$ = 2.0%
Red spot
Channel 187
$R_{187}(x|y)$ = 1.2%
Black eyes
d Hierarchical concept composition
features.24
features.26
features.28
Output
Channel 263: colourful feathers (1%)
Channel 407: colourful threads (1%)
Channel 495: wood (horizontal) (2%)
Channel 118: brown, knobby (2%)
4%
5%
3%
12%
Channel 506: colourful, bushy feathers (4%)
Channel 51: crossed bars (3%)
4%
3%
Channel 102: animal on branch (100%)
5%
Bee Eater
Reference samples (most relevant)
(Multi)-conditional heatmap
Relevance flow

LIME (Local Interpretable Model-agnostic Explanations):

Function: LIME explains predictions of any classifier in an interpretable manner.

Method: Generates perturbations of input data, observes changes in predictions, and creates a local linear model around the prediction to identify key features.

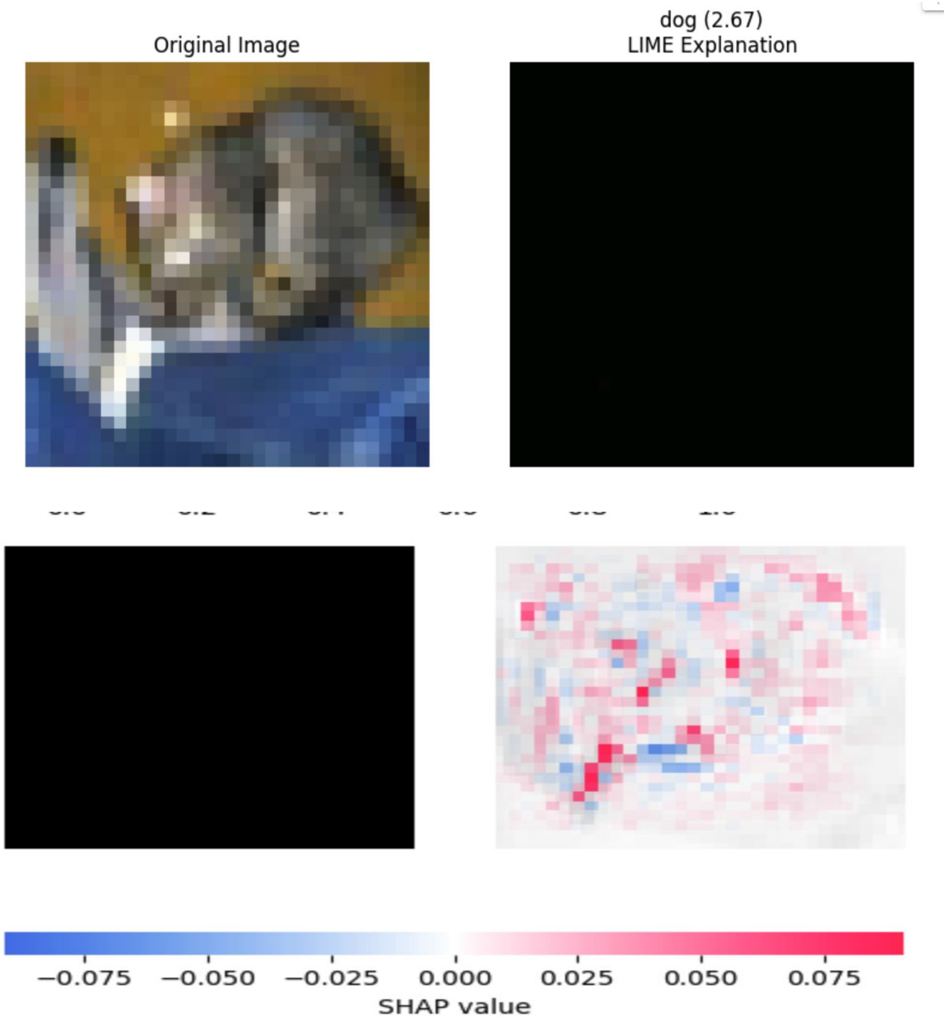
Output: Produces a visualization highlighting parts of the input (like pixels in an image) most influential in the model's decision.

SHAP (SHapley Additive exPlanations):

Function: Breaks down a prediction to show the impact of each feature.

Method: Uses Shapley values to measure the contribution of each feature, considering all possible combinations of features.

Output: Provides detailed insights into how each feature contributes to the model's output, often visualized through plots that show feature impact on model output.



Objective: Develop a CNN model using TensorFlow and Keras to classify images from the CIFAR-10 dataset.

Model Architecture: The CNN consists of convolutional layers, pooling layers, and dense layers.

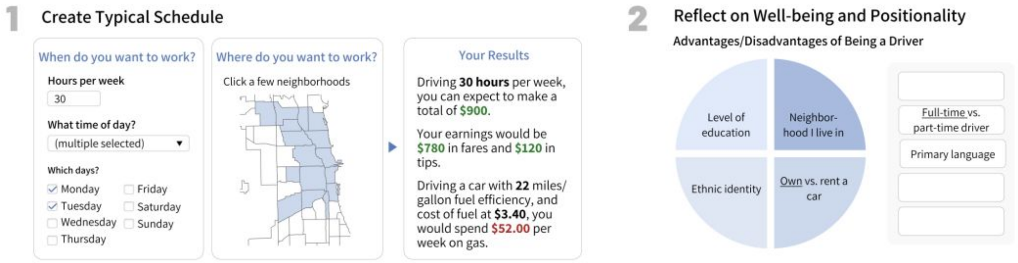
Training: The model is trained on the CIFAR-10 dataset to categorize images into ten classes.

Co-Design

Co-Design Activities and Data Probes address a variety of these problems, and also meaningfully advance design.

Can offer explanation in data trends, design recommendations, prototype evaluations, data on user mental models, and more - all within the same interview session!

Overview of the Co-Design Procedure



1 Create Typical Schedule

When do you want to work?

Hours per week: 30

What time of day? (multiple selected) ▾

Which days?

☑ Monday ☐ Friday
☑ Tuesday ☐ Saturday
☐ Wednesday ☐ Sunday
☐ Thursday

Where do you want to work?

Click a few neighborhoods

Your Results

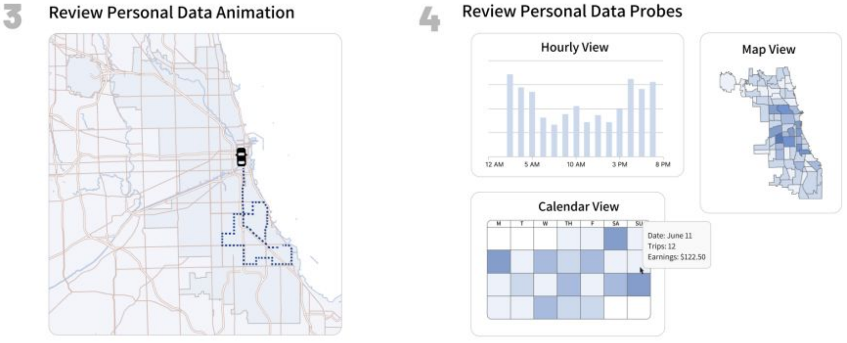
Driving **30 hours** per week, you can expect to make a total of $900.

Your earnings would be $780 in fares and $120 in tips.

Driving a car with **22** miles/gallon fuel efficiency, and cost of fuel at **$3.40**, you would spend **$52.00** per week on gas.

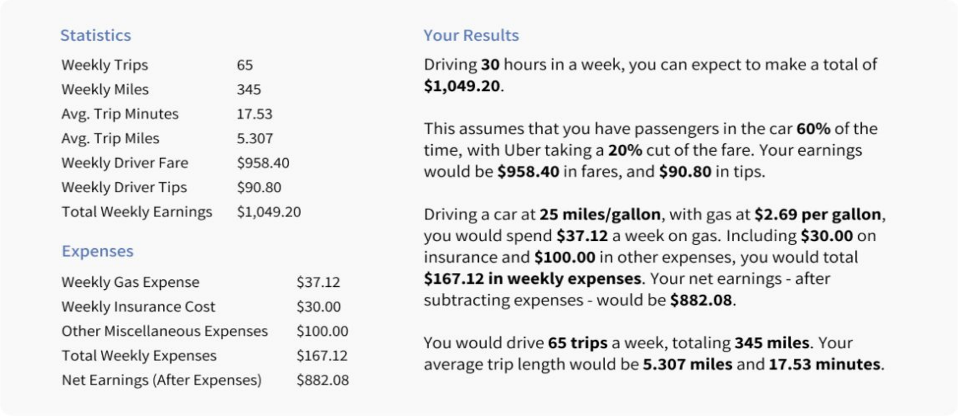
2 Reflect on Well-being and Positionality

Advantages/Disadvantages of Being a Driver



3 Review Personal Data Animation

4 Review Personal Data Probes



Statistics

Weekly Trips	65
Weekly Miles	345
Avg. Trip Minutes	17.53
Avg. Trip Miles	5.307
Weekly Driver Fare	$958.40
Weekly Driver Tips	$90.80
Total Weekly Earnings	$1,049.20

Expenses

Weekly Gas Expense	$37.12
Weekly Insurance Cost	$30.00
Other Miscellaneous Expenses	$100.00
Total Weekly Expenses	$167.12
Net Earnings (After Expenses)	$882.08

Your Results

Driving **30** hours in a week, you can expect to make a total of **$1,049.20**.

This assumes that you have passengers in the car **60%** of the time, with Uber taking a **20%** cut of the fare. Your earnings would be **$958.40** in fares, and **$90.80** in tips.

Driving a car at **25 miles/gallon**, with gas at **$2.69 per gallon**, you would spend **$37.12** a week on gas. Including **$30.00** on insurance and **$100.00** in other expenses, you would total **$167.12 in weekly expenses**. Your net earnings - after subtracting expenses - would be **$882.08**.

You would drive **65 trips** a week, totaling **345 miles**. Your average trip length would be **5.307 miles** and **17.53 minutes**.

Figure 2: Work Planner data probe that participants interacted with to view predictions of their schedules and surface design considerations.

Thank you