XAI FRAMEWORK OF REASONED EXPLANATIONS

Numerous novel explanation techniques have been developed for explainable AI (XAI). How can developers choose which technique to implement for various users and use cases? We present the XAI Framework of Reasoned Explanations to help guide how to choose different XAI feature based on user goals and human reasoning methods and biases.

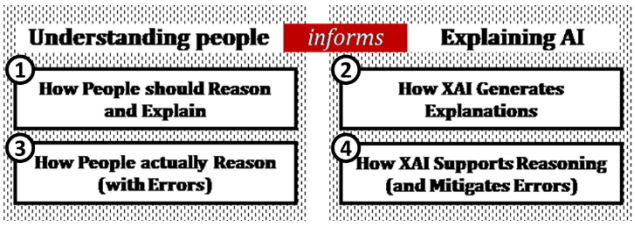

Figure 1. Overview of user-centric XAI framework.

The framework describes how people reason rationally (Figure 1, ①) and heuristically but subject to cognitive biases ③, how XAI facilities do support specific rational reasoning processes ②, and can be designed to target decision errors ④. The framework identifies pathways between human reasoning and XAI facilities that can help organize explanations and identify gaps to develop new explanations given an unmet reasoning need.

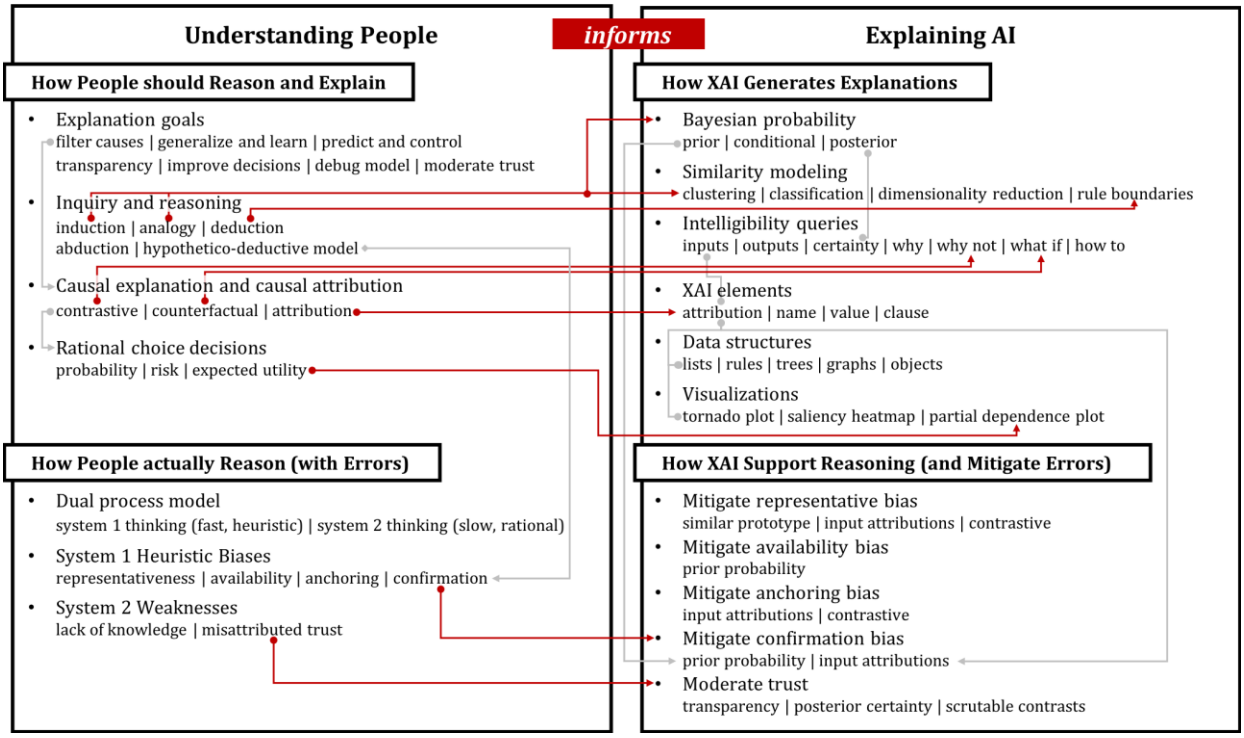

Figure 2. Conceptual framework for Reasoned Explanations that describes how human reasoning processes (left) informs XAI techniques (right). Points describe different theories of reasoning, XAI techniques, and strategies for designing XAI. Arrows indicate pathway connections: red arrows for how theories of human reasoning inform XAI features, and grey for inter-relations between different reasoning processes and associations between XAI features. Only some example pathways are shown. For example, hypothetico-deductive reasoning can be interfered by System 1 thinking and cause confirmation bias (grey arrow). Confirmation bias can be mitigated (follow the red line) by presenting information about the prior probability or input attributions. Next, we can see that input attributions can be implemented as lists and visualized using tornado plots (follow the grey line). We identified many pathways but show only a few for brevity.

Procedure to use Framework

XAI application developers can use the framework as follows:

1. Consider the user's reasoning goals ① and biases ③ for their respective apps. This can be informed through literature review, ethnography, participatory design, etc.
2. Next, identify which explanations help reasoning goals ② or reduce cognitive biases ④ using pathways in the framework (Figure 2, red arrows).
3. Finally, integrate these XAI facilities to create explainable UIs.

Furthermore, XAI researchers can extend the framework by

1) Examining new XAI facilities ② to understand how they were inspired by, dependent on, or built from reasoning theories ① and
2) Identifying common biases and reasoning errors ③ related to reasoning theories ① and then identifying appropriate mitigation strategies ② to select specific XAI facilities.

For example, Informal Logic could be integrated into reasoning theories ① and informal fallacies into reasoning errors ③.

Example Application: Medical Diagnosis

Figure 3. Screenshot of the AI-driven medical diagnosis tool with explanation sketches showing a patient with high predicted risk of acute myocardial infarction (AMI), heart disease, diabetes with complications, shock, etc.

Explanations include

- (top left) feature value time series,
- (top right) class attribution of predicted disease risk,
- (middle right) feature attribution by vitals, and
- (bottom) counterfactual rules indicating key rules for each prediction.

Interpretation: e.g., explanations suggest that the AI thinks that the patient has shock because of low oxygen saturation and blood pressure.

Glossary (some terms)

- **Deductive reasoning**: process of reasoning from premises to a conclusion.
- **Inductive reasoning**: from a single observation or instance to a probable explanation or generalization.
- **Abductive reasoning**: from an observation to the most likely explanation. Also known as “inference to the best explanation”.
- **Hypothetico-Deductive model**: steps of inquiry as (1) observe and identify a new problem, (2) form a hypothesis as induction from observations, (3) deduce consequent predictions from the hypotheses, and (4) test (run experiments) or look for (or fail to find) further observations that falsify hypotheses.
- **Analogical reasoning:** from one instance to another. It is a weak form of induction since only one instance is considered.
- **Causal attribution:** articulation of internal or external factors that could be attributed to influence the outcome.
- **Causal explanation:** focused on *selected* causes relevant to interpreting the observation with respect to existing knowledge. This requires that the explanation be ***contrastive*** between a *fact* (what happened) and a *foil* (what is expected or plausible to happen). Users can ask *why not* to understand why a foil did not happen. The selected subset of causes thus provides a ***counterfactual*** explanation of what needs to change for the alternative outcome to happen. This helps users to identify causes, on the principle that manipulating a cause will change the effect.
- **System 1** (of Kahneman’s dual process model) thinking is fast, intuitive, and low effort, where the person employs cognitive heuristics to make quick decisions, but this can lead to cognitive biases. Example effects: ***representativeness*** heuristic to infer the similarity of an event or item to a previously observed item; ***availability bias*** can occur when decision makers are unfamiliar with how often a particular outcome happens; ***confirmation bias***, where the decision maker tends to look for confirming evidence of a wrongly-chosen hypothesis.
- **System 2** thinking is slow, analytical, and high effort, where the person employs rational reasoning processes. Even without cognitive biases, System 2 thinking may still fail due to using a ***miscalibrated*** tool or a ***lack of domain knowledge***.

Further Reading

Wang, D., Yang, Q., Abdul, A., Lim, B. Y. 2019. **Designing Theory-Driven User-Centric Explainable AI.** In *Proceedings of the international Conference on Human Factors in Computing Systems. CHI ’19*.