Formal Explanations in Artificial Intelligence

ESSAI 2025

Julien Girard-Satabin
Zakaria Chihani
Dorin Doncenco

CEA LIST

2025-07-02

This work was supported by the French Agence Nationale de la Recherche (ANR) through SAIF (ANR-23-PEIA-0006) and DeepGreen (ANR-23-DEGR-0001) as part of the France 2030 programme.

About lecturer

Figure 1: Questions may be asked during the lecture!

- BSc in Computer Science & Engineering
- MSc in Artificial Intelligence (AI)
- Interested in applied eXplainable AI (XAI)
- PhD Student on Formal and Actionable XAI
- Likes applying the handshake protocol to human conversations

Session goal

- Understand importance of eXplainability in AI systems
- Discuss what makes an explanation
- Learn limitations of common XAI methods
- Address these limitations with Formal XAI (FXAI)
- Discover new FXAI concepts
- Practical results with FXAI

Motivation

Omnipresence of automated systems

Omnipresence of automated systems

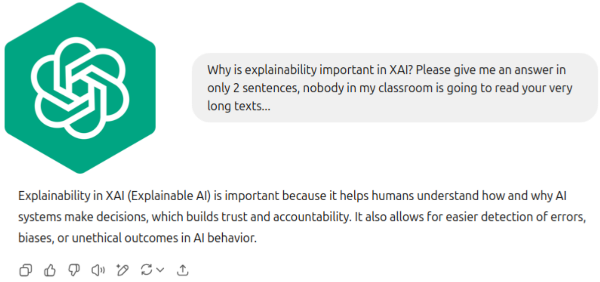


Figure 2: LLMs are large models that have seen quick adoption within society.

Omnipresence of automated systems

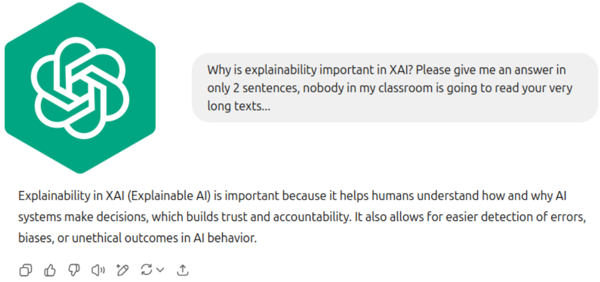


Figure 2: LLMs are large models that have seen quick adoption within society.

Figure 3: AI might be adopted in education.

Omnipresence of automated systems

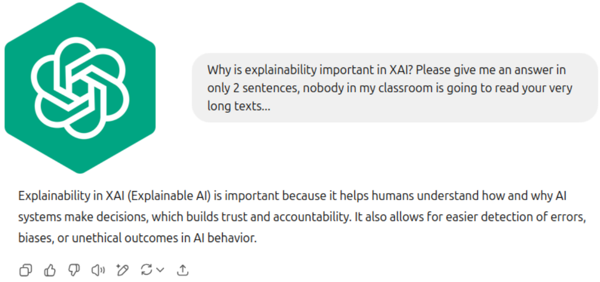


Figure 2: LLMs are large models that have seen quick adoption within society.

Figure 3: AI might be adopted in education.

Figure 4: High stakes for avoiding harm from autonomous vehicles.

Omnipresence of automated systems

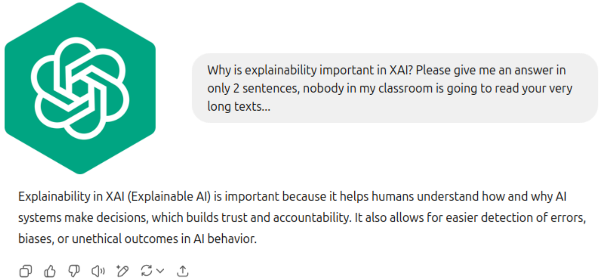


Figure 2: LLMs are large models that have seen quick adoption within society.

Figure 3: AI might be adopted in education.

Figure 4: High stakes for avoiding harm from autonomous vehicles.

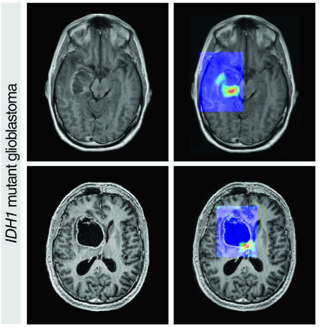


Figure 5: Automatic tumor detection has the potential to save lives.

Omnipresence of automated systems

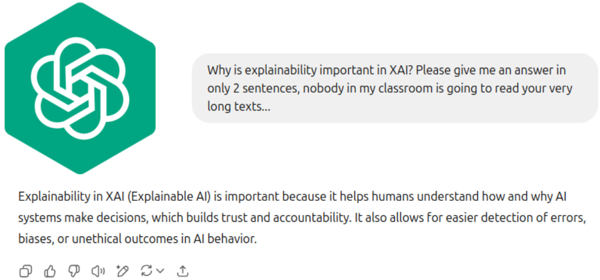


Figure 2: LLMs are large models that have seen quick adoption within society.

Figure 3: AI might be adopted in education.

Figure 4: High stakes for avoiding harm from autonomous vehicles.

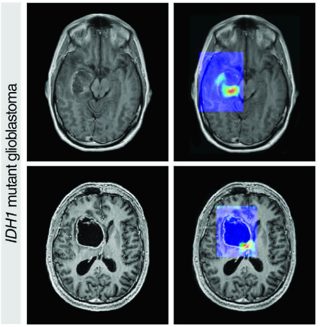
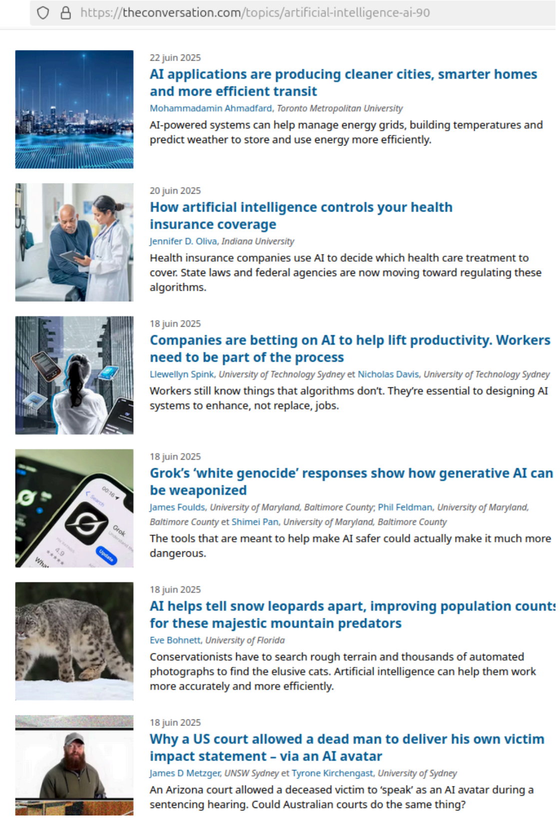


Figure 5: Automatic tumor detection has the potential to save lives.



https://theconversation.com/topics/artificial-intelligence-ai-90

22 juin 2025
AI applications are producing cleaner cities, smarter homes and more efficient transit
Mohammadamin Ahmadfard, *Toronto Metropolitan University*
AI-powered systems can help manage energy grids, building temperatures and predict weather to store and use energy more efficiently.

20 juin 2025
How artificial intelligence controls your health insurance coverage
Jennifer D. Oliva, *Indiana University*
Health insurance companies use AI to decide which health care treatment to cover. State laws and federal agencies are now moving toward regulating these algorithms.

18 juin 2025
Companies are betting on AI to help lift productivity. Workers need to be part of the process
Llewellyn Spink, *University of Technology Sydney* et Nicholas Davis, *University of Technology Sydney*
Workers still know things that algorithms don't. They're essential to designing AI systems to enhance, not replace, jobs.

18 juin 2025
Grok's 'white genocide' responses show how generative AI can be weaponized
James Foulds, *University of Maryland, Baltimore County*; Phil Feldman, *University of Maryland, Baltimore County* et Shimei Pan, *University of Maryland, Baltimore County*
The tools that are meant to help make AI safer could actually make it much more dangerous.

18 juin 2025
AI helps tell snow leopards apart, improving population counts for these majestic mountain predators
Eve Bohnett, *University of Florida*
Conservationists have to search rough terrain and thousands of automated photographs to find the elusive cats. Artificial intelligence can help them work more accurately and more efficiently.

18 juin 2025
Why a US court allowed a dead man to deliver his own victim impact statement – via an AI avatar
James D Metzger, *UNSW Sydney* et Tyrone Kirchengast, *University of Sydney*
An Arizona court allowed a deceased victim to 'speak' as an AI avatar during a sentencing hearing. Could Australian courts do the same thing?

Figure 6: AI is a discussion point on many important topics.

The impact of automated decision systems

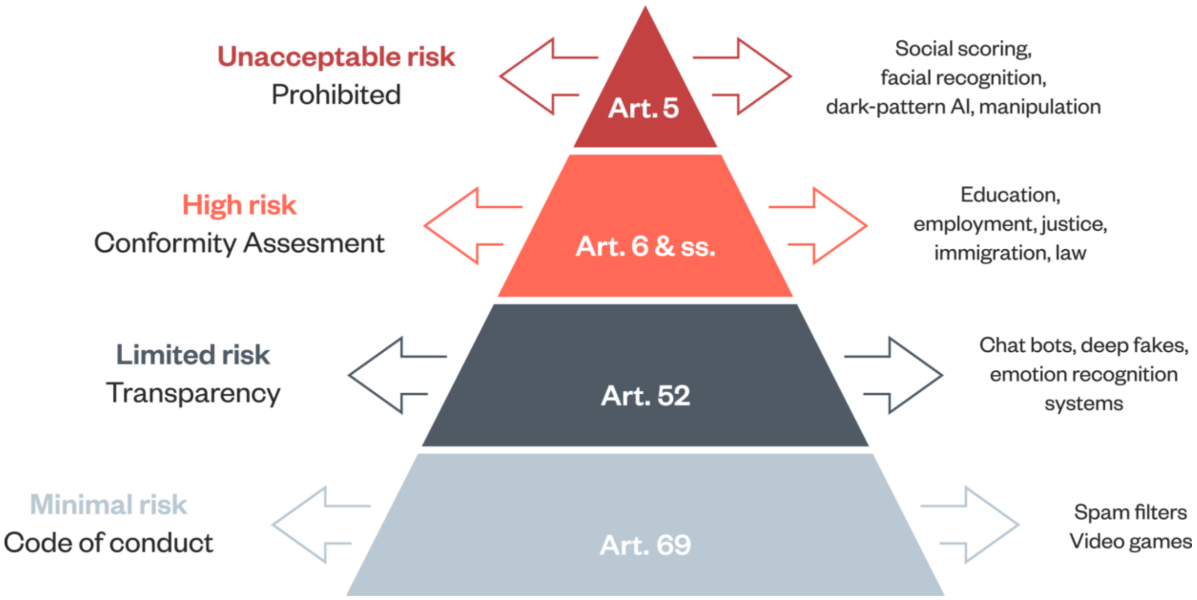


Figure 7: Risk hierarchy of AI use-cases[1]. Various institutions have differing definitions.

[1] https://www.adalovelaceinstitute.org/resource/eu-ai-act-explainer/

[2] https://en.wikipedia.org/wiki/Right_to_explanation

The impact of automated decision systems

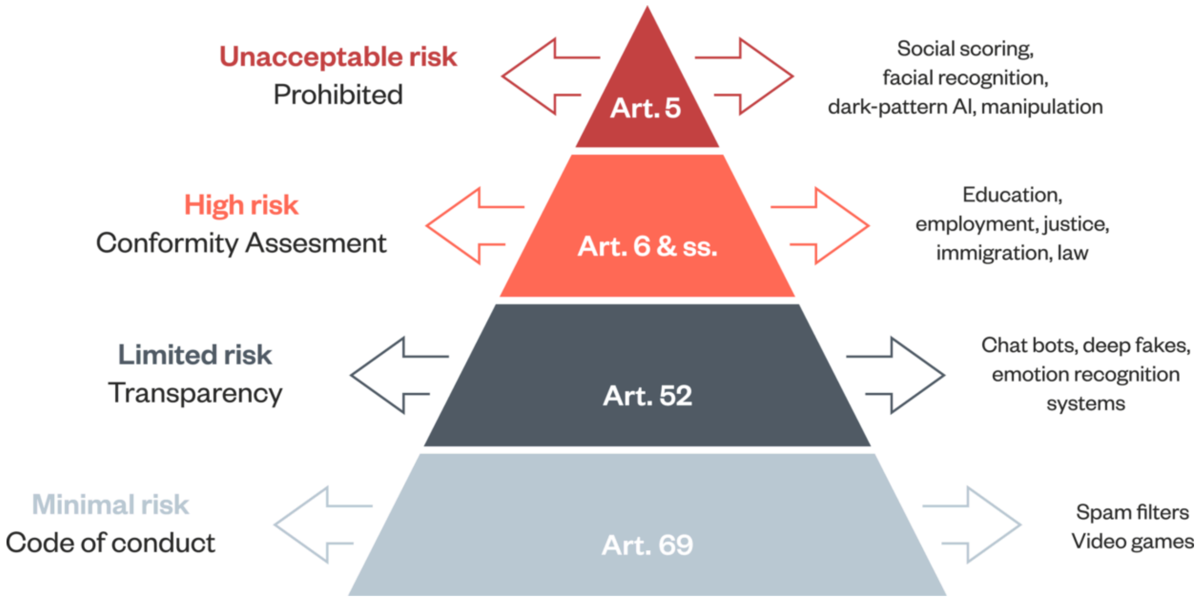


Figure 7: Risk hierarchy of AI use-cases[1]. Various institutions have differing definitions.

Regulating high risks involves:

- transparency and information to users
- right to explanation[2]
- ensuring the safety of these systems

[1] https://www.adalovelaceinstitute.org/resource/eu-ai-act-explainer/

[2] https://en.wikipedia.org/wiki/Right_to_explanation

Explanations | Introduction

Proof vs Explanations

- The concept of explanation is illusive
- Not easy to explain something

Proof vs Explanations

- The concept of explanation is illusive
- Not easy to explain something

Let’s demonstrate with an example:

Proof vs Explanations

- The concept of explanation is illusive
- Not easy to explain something

Let's demonstrate with an example:

$$12345679 \times 36 = 444444444$$

Proof vs Explanations

- The concept of explanation is illusive
- Not easy to explain something

Let's demonstrate with an example:

$$12345679 \times 36 = 444444444$$

Why is this result particular such that it consists only of the digit 4?

First attempt - Demonstration

$$
\begin{array}{r}
12345679 \\
\times \quad 36 \\
\hline
\end{array}
$$

First attempt - Demonstration

$$
\begin{array}{r}
12345679 \\
\times \quad 36 \\
\hline
74074074
\end{array}
$$

First attempt - Demonstration

$$
\begin{array}{r}
12345679 \\
\times \qquad\quad 36 \\
\hline
74074074 \\
+37037037
\end{array}
$$

First attempt - Demonstration

$$
\begin{array}{r}
12345679 \\
\times \qquad\quad 36 \\
\hline
74074074 \\
+37037037 \\
\hline
\end{array}
$$

First attempt - Demonstration

$$
\begin{array}{r}
12345679 \\
\times \qquad\quad 36 \\
\hline
74074074 \\
+37037037 \\
\hline
444444444
\end{array}
$$

First attempt - Demonstration

$$
\begin{array}{r}
12345679 \\
\times \qquad\quad 36 \\
\hline
74074074 \\
+37037037 \\
\hline
444444444
\end{array}
$$

- What is your take on such an explanation?

First attempt - Demonstration

$$
\begin{array}{r}
12345679 \\
\times \qquad\quad 36 \\
\hline
74074074 \\
+37037037 \\
\hline
444444444
\end{array}
$$

- What is your take on such an explanation?
- What is lacking about the explanation?

First attempt – Demonstration

$$
\begin{array}{r}
12345679 \\
\times \qquad\quad 36 \\
\hline
74074074 \\
+37037037 \\
\hline
444444444
\end{array}
$$

- What is your take on such an explanation?
- What is lacking about the explanation?
- We have proven the truthfulness (good)
- It doesn't help understand the result (bad)

First attempt - Demonstration

$$
\begin{array}{r}
12345679 \\
\times \qquad\quad 36 \\
\hline
74074074 \\
+37037037 \\
\hline
444444444
\end{array}
$$

- What is your take on such an explanation?
- What is lacking about the explanation?
- We have proven the truthfulness (good)
- It doesn't help understand the result (bad)
- Additional questions:
- Why does this result work for 444,444,444?
- Can we make it work for 555,555,555?

Second attempt - Explanation #1

$$444444444 = 12345679 \times 36$$
$$444444444 = 111111111 \times 4$$
$$111111111 = 12345679 \times 9$$

- This explanation allows us to generalize the example;

Second attempt - Explanation #1

$$
\begin{aligned}
444444444 &= 12345679 \times 36 \\
444444444 &= 111111111 \times 4 \\
111111111 &= 12345679 \times 9
\end{aligned}
$$

- This explanation allows us to generalize the example;
- What other questions are left out?

Second attempt - Explanation #1

$$
\begin{aligned}
444444444 &= 12345679 \times 36 \\
444444444 &= 111111111 \times 4 \\
111111111 &= 12345679 \times 9
\end{aligned}
$$

- This explanation allows us to generalize the example;
- What other questions are left out?
- We still don't know why 12345679 is interesting here
- Why does 111,111,111 have 9 digits?
- Why is the number 8 missing?

Third attempt - Explanation #2

$$111111111 \,|\, 9$$

Third attempt - Explanation #2

$$\begin{array}{l|l} 111111111 & 9 \\ \hline & 1 \end{array}$$

Third attempt – Explanation #2

$$\begin{array}{l|l} 111111111 & 9 \\ \;\;21 & \overline{} \\ & 12 \end{array}$$

Third attempt – Explanation #2

$$
\begin{array}{l|l}
111111111 & 9 \\ \cline{2-2}
\ \ 21 & 123 \\
\ \ \ \ 31 &
\end{array}
$$

Third attempt - Explanation #2

$$
\begin{array}{l|l}
111111111 & 9 \\
\quad 21 & \overline{} \\
\quad\; 31 & 1234 \\
\quad\;\; 41 &
\end{array}
$$

Third attempt – Explanation #2

$$\mathbf{11}\,11\,11111 \mid 9$$

$$\mathbf{21}$$

$$\mathbf{31} \qquad \mathbf{1234}$$

$$\mathbf{41}$$

Third attempt - Explanation #2

$$\mathbf{11}\,11\,11111|9$$

$$\mathbf{21}$$

$$\mathbf{31} \qquad \mathbf{1234}$$

$$\mathbf{41}$$

$$10 \times n + 1 = (9 \times n) + (n + 1)$$

Third attempt - Explanation #2

$$\mathbf{11}\,11\,11111 \,|\, 9$$

$$\mathbf{21}$$

$$\mathbf{31} \qquad \overline{\mathbf{1234}}$$

$$\mathbf{41}$$

$$10 \times n + 1 = (9 \times n) + (n + 1)$$

- left side: tens and ones
- right side: quotient + remainder

Third attempt - Explanation #2

$\mathbf{11}11\,11111|9$

$\mathbf{21}$ $\quad$ ————

$\mathbf{31}$ $\quad \mathbf{1234}$

$\mathbf{41}$

$$10 \times n + 1 = (9 \times n) + (n + 1)$$

- left side: tens and ones
- right side: quotient + remainder
- At $n = 8$, we get the remainder 0

Third attempt - Explanation #2

$$
\begin{array}{l}
111111111 | 9 \\
\quad 21 \qquad\qquad \overline{} \\
\quad\; 31 \qquad\quad 1234567 \\
\quad\;\; 41 \\
\quad\;\;\; 51 \\
\quad\;\;\;\; 61 \\
\quad\;\;\;\;\; 71 \\
\quad\;\;\;\;\;\; 81
\end{array}
$$

$$10 \times n + 1 = (9 \times n) + (n + 1)$$

- left side: tens and ones
- right side: quotient + remainder
- At $n = 8$, we get the remainder 0

Third attempt - Explanation #2

$$
\begin{array}{l|l}
111111111 & 9 \\
\quad 21 & \overline{} \\
\quad\; 31 & 12345679 \\
\quad\quad 41 & \\
\quad\quad\; 51 & \\
\quad\quad\quad 61 & \\
\quad\quad\quad\; 71 & \\
\quad\quad\quad\quad 81 & \\
\quad\quad\quad\quad\; 0 &
\end{array}
$$

$$10 \times n + 1 = (9 \times n) + (n + 1)$$

- left side: tens and ones
- right side: quotient + remainder
- At $n = 8$, we get the remainder 0

Takeaways

- There are multiple ways of explaining something
- Proving a statement does not necessarily lead to an insight into the problem

[1] https://www.youtube.com/watch?v=6j8Vwbss038 (by Gilles Dowek, in French)

Takeaways

- There are multiple ways of explaining something
- Proving a statement does not necessarily lead to an insight into the problem
- Exercise source[1]

[1] https://www.youtube.com/watch?v=6j8Vwbss038 (by Gilles Dowek, in French)

Interactive Time

- What makes an explanation?
- *Skip*

Interactive Time

- What makes an explanation?
- Case study
- *I have gotten funds for a new scholarship program, and I am looking for **good students** who would benefit from it. Conveniently, I have **decided** that I am a **good student** who should benefit from this scholarship.*

Interactive Time

- What makes an explanation?
- Case study
- *I have gotten funds for a new scholarship program, and I am looking for **good students** who would benefit from it. Conveniently, I have **decided** that I am a **good student** who should benefit from this scholarship.*
- Should this decision be explained?
- What kind of explanations would you like?

General consensus

- An explanation is the product of interaction between a stakeholder and a system, with the goal of the stakeholder extracting knowledge about the system, and its decision.[1][2]

[1]"Metrics for explainable AI: Challenges and prospects", Hoffman et al.

[2]"Algorithms and evaluation metrics for improving trust in machine learning : application to visual object recognition", Romain Xu-Darme

[3]*Interpretable Machine Learning* by Christoph Molnar

[4]"Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems", Tomsett et al.

General consensus

- An explanation is the product of interaction between a stakeholder and a system, with the goal of the stakeholder extracting knowledge about the system, and its decision.[1][2]
- Explanations are not the goal, but a way of achieving other goals[3][4]:

[1] "Metrics for explainable AI: Challenges and prospects", Hoffman et al.

[2] "Algorithms and evaluation metrics for improving trust in machine learning : application to visual object recognition", Romain Xu-Darme

[3] *Interpretable Machine Learning* by Christoph Molnar

[4] "Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems", Tomsett et al.

General consensus

- An explanation is the product of interaction between a stakeholder and a system, with the goal of the stakeholder extracting knowledge about the system, and its decision.[1,2]
- Explanations are not the goal, but a way of achieving other goals[3,4]:
  - Discover insights about the problem studied
  - Improve/debug models
  - Justify predictions (and models) to stakeholders

[1] "Metrics for explainable AI: Challenges and prospects", Hoffman et al.

[2] "Algorithms and evaluation metrics for improving trust in machine learning : application to visual object recognition", Romain Xu-Darme

[3] *Interpretable Machine Learning* by Christoph Molnar

[4] "Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems", Tomsett et al.

Stakeholders for explanations

- developers
- operators
- executors
- auditors
- data subjects
- decision subjects

Interesting properties of explanations

	Co-12 Property	Description
Content	**Correctness**	Describes how faithful the explanation is w.r.t. the black box.
		Key idea: Nothing but the truth
	Completeness	Describes how much of the black box behavior is described in the explanation.
		Key idea: The whole truth
	Consistency	Describes how deterministic and implementation-invariant the explanation method is.
		Key idea: Identical inputs should have identical explanations
	Continuity	Describes how continuous and generalizable the explanation function is.
		Key idea: Similar inputs should have similar explanations
	Contrastivity	Describes how discriminative the explanation is w.r.t. other events or targets.
		Key idea: Answers "why not?" or "what if?" questions
	Covariate complexity	Describes how complex the (interactions of) features in the explanation are.
		Key idea: Human-understandable concepts in the explanation
Presentation	**Compactness**	Describes the size of the explanation.
		Key idea: Less is more
	Composition	Describes the presentation format and organization of the explanation.
		Key idea: *How* something is explained
	Confidence	Describes the presence and accuracy of probability information in the explanation.
		Key idea: Confidence measure of the explanation or model output
User	**Context**	Describes how relevant the explanation is to the user and their needs.
		Key idea: How much does the explanation matter in practice?
	Coherence	Describes how accordant the explanation is with prior knowledge and beliefs.
		Key idea: Plausibility or reasonableness to users
	Controllability	Describes how interactive or controllable an explanation is for a user.
		Key idea: Can the user influence the explanation?

Figure 34: Properties of explanations, as identified in[1]

[1]"From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI" by Nauta et al.

Interesting properties of explanations

	Co-12 Property	Description
Content	**Correctness**	Describes how faithful the explanation is w.r.t. the black box.
		Key idea: Nothing but the truth
	Completeness	Describes how much of the black box behavior is described in the explanation.
		Key idea: The whole truth
	Consistency	Describes how deterministic and implementation-invariant the explanation method is.
		Key idea: Identical inputs should have identical explanations
	Continuity	Describes how continuous and generalizable the explanation function is.
		Key idea: Similar inputs should have similar explanations
	Contrastivity	Describes how discriminative the explanation is w.r.t. other events or targets.
		Key idea: Answers "why not?" or "what if?" questions
	Covariate complexity	Describes how complex the (interactions of) features in the explanation are.
		Key idea: Human-understandable concepts in the explanation
Presentation	**Compactness**	Describes the size of the explanation.
		Key idea: Less is more
	Composition	Describes the presentation format and organization of the explanation.
		Key idea: *How* something is explained
	Confidence	Describes the presence and accuracy of probability information in the explanation.
		Key idea: Confidence measure of the explanation or model output
User	**Context**	Describes how relevant the explanation is to the user and their needs.
		Key idea: How much does the explanation matter in practice?
	Coherence	Describes how accordant the explanation is with prior knowledge and beliefs.
		Key idea: Plausibility or reasonableness to users
	Controllability	Describes how interactive or controllable an explanation is for a user.
		Key idea: Can the user influence the explanation?

Figure 34: Properties of explanations, as identified in[1]

Figure 35: Poll: Which properties do you view as most important? Link: https://whale5.noiraudes.net/polls/337317df-8895-402b-9ddb-0dea7564a1e9

[1]"From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI" by Nauta et al.

Classical vs Formal XAI

Classical XAI

- Problem: system opacity[1]

[1]Further reading: "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead" by Cynthia Rudin

[2]https://christophm.github.io/interpretable-ml-book/overview.html

[3]https://dept.utc2.edu.vn/bomoncntt/doi-tac/decision-trees-explained-with-a-practical-example-57.html

Classical XAI

- Problem: system opacity[1]

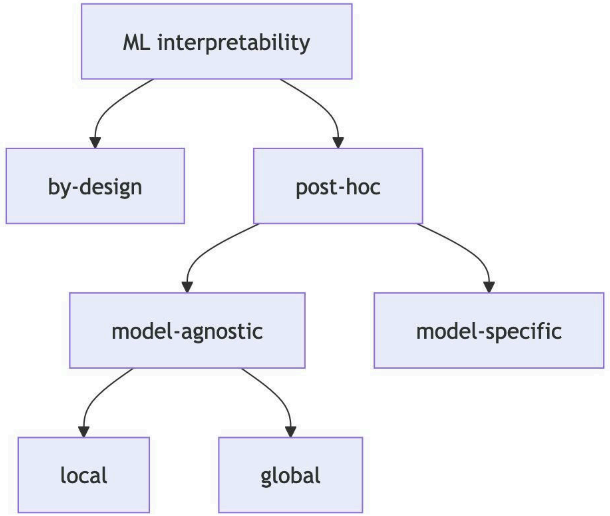


Figure 36: Taxonomy of interpretability methods[2]

[1]Further reading: "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead" by Cynthia Rudin

[2]https://christophm.github.io/interpretable-ml-book/overview.html

[3]https://dept.utc2.edu.vn/bomoncntt/doi-tac/decision-trees-explained-with-a-practical-example-57.html

Classical XAI

- Problem: system opacity[1]

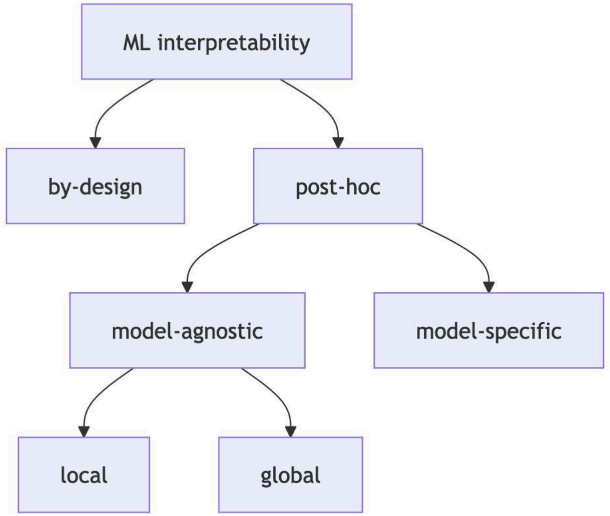


Figure 36: Taxonomy of interpretability methods[2]

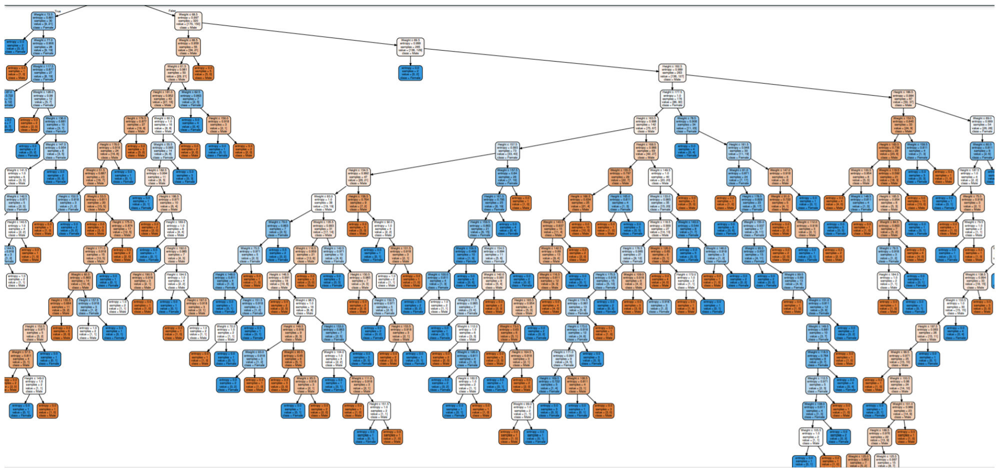


Figure 37: Deep decision tree.[3]

[1]Further reading: "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead" by Cynthia Rudin

[2]https://christophm.github.io/interpretable-ml-book/overview.html

[3]https://dept.utc2.edu.vn/bomoncntt/doi-tac/decision-trees-explained-with-a-practical-example-57.html

What classical XAI had promised to us:

Classical XAI aims to[1] :

- help us improve our model
- justify answers to various stakeholders
- discover insights about the problem

[1]https://christophm.github.io/interpretable-ml-book/goals.html

AI risk scenarios (again)

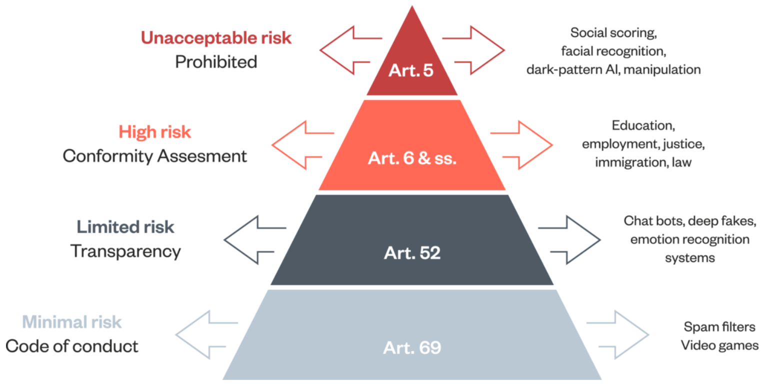


Figure 38: Reminder of AI risk hierarchy.

It is untolerable to:

- have systems that cannot be understood
- have mistakes in explanations

Limitations of traditional XAI

The explanations should be:

- compact (no redundant features)
- faithful to the model

[1]"Disproving XAI Myths with Formal Methods – Initial Results" by Joao Marques-Silva

Limitations of traditional XAI

The explanations should be:

- compact (no redundant features)
- faithful to the model

Limitations of XAI

- Not respecting these requirements

[1] "Disproving XAI Myths with Formal Methods – Initial Results" by Joao Marques-Silva

Limitations of traditional XAI

The explanations should be:
- compact (no redundant features)
- faithful to the model

Limitations of XAI
- Not respecting these requirements
- many others (but we will focus on the mentioned ones)

[1] "Disproving XAI Myths with Formal Methods – Initial Results" by Joao Marques-Silva

Limitations of traditional XAI

The explanations should be:

- compact (no redundant features)
- faithful to the model

Limitations of XAI

- Not respecting these requirements
- many others (but we will focus on the mentioned ones)

Case	Instance	Relevant	Irrelevant	Shapley values	Justification
I4	$((0,0,1,1),0)$	$1,2,4$	3	$\mathsf{Sv}(1) = -0.13$ $\mathsf{Sv}(2) = 0.33$ $\mathsf{Sv}(3) = 0.08$ $\mathsf{Sv}(4) = 0.00$	$\mathsf{Irrelevant}(3) \wedge \mathsf{Sv}(3) \neq 0 \wedge$ $\mathsf{Relevant}(4) \wedge \mathsf{Sv}(4) = 0$
I5	$((1,1,1,1),0)$	$1,2,3$	4	$\mathsf{Sv}(1) = -0.12$ $\mathsf{Sv}(2) = -0.12$ $\mathsf{Sv}(3) = -0.12$ $\mathsf{Sv}(4) = 0.17$	$\mathsf{Irrelevant}(4) \wedge$ $\forall (j \in \{1,2,3\}).\lvert\mathsf{Sv}(j)\rvert < \mathsf{Sv}(4)\rvert$

TABLE IV: Examples of issues with Shapley for explainability for boolean classifiers of Figure 3

Figure 39: Shapley Values can return non-zero scores for irrelevant features, and zero scores for relevant ones.[1]

[1] "Disproving XAI Myths with Formal Methods – Initial Results" by Joao Marques-Silva

Advantages of Formal XAI

Formal XAI (FXAI) is able to:

- compute compact explanations
- compute model-faithful explanations
- validate the robustness of a model wrt. requirements

Formal XAI | Building Safe AI Systems

Toy problem

- Let’s formalize a system that detects mammals!

Figure 40: Platypodes are weird mammals.

Toy problem

- Let’s formalize a system that detects mammals!
- (disclaimer: not a biologist)

Figure 42: Platypodes are weird mammals.

Classification problem

Consider a classification instance,

- X: Feature space
- C: classification space
- x, c : an instance of a vector of values in X and corresponding class in C
- f: classification function, eg. $f(x) = c$

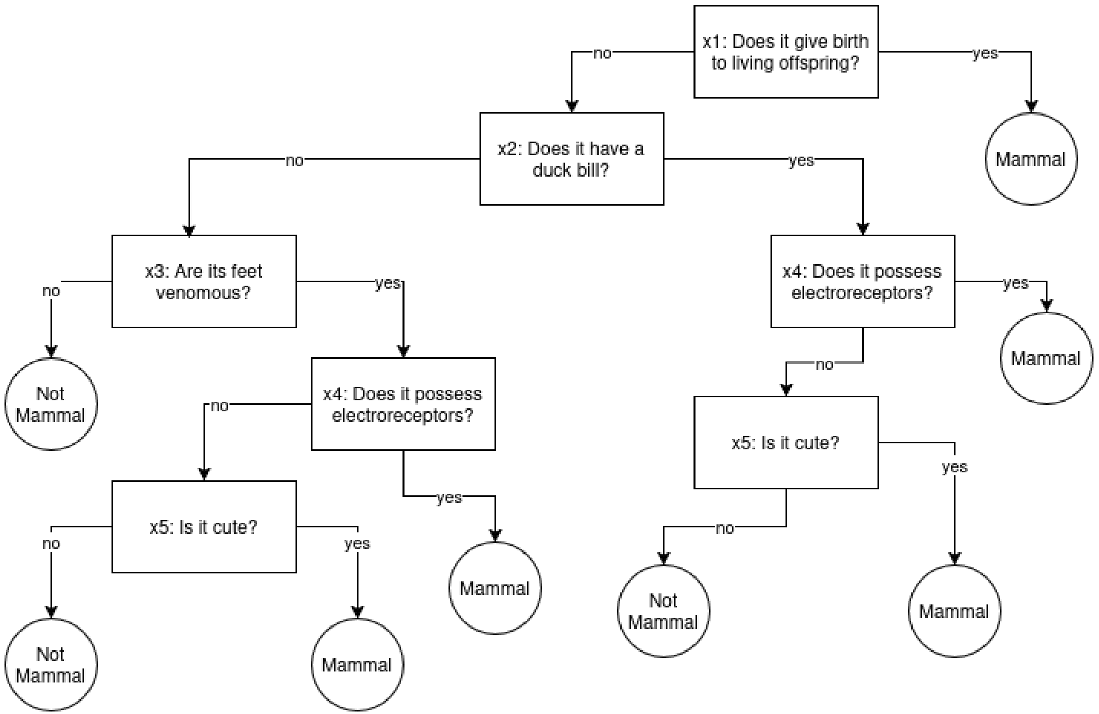


Figure 43: Adapted decision tree example[1]

[1]"Disproving XAI Myths with Formal Methods – Initial Results", Joao Marques-Silva

Classification problem

- $X = \{X_i \in \{0,1\}, \forall i \in \{1,2,3,4,5\}\}$
- $C = \{0,1\}$
- $x = \{1,0,1,0,1\}$
- $c = 1$

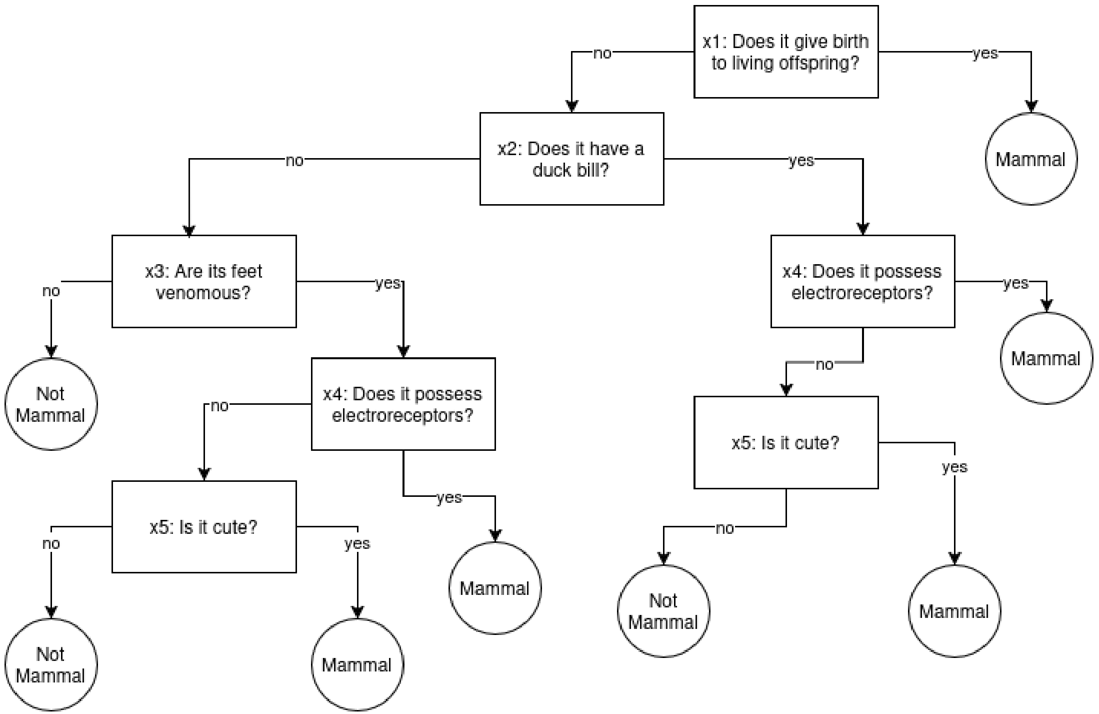


Figure 47: Adapted decision tree example[1]

[1]"Disproving XAI Myths with Formal Methods – Initial Results", Joao Marques-Silva

Classification problem

- $X = \{X_i \in \{0,1\}, \forall i \in \{1,2,3,4,5\}\}$
- $C = \{0,1\}$
- $x = \{1,0,1,0,1\}$
- $c = 1$

Why did the model classify $f(x) = 1$?

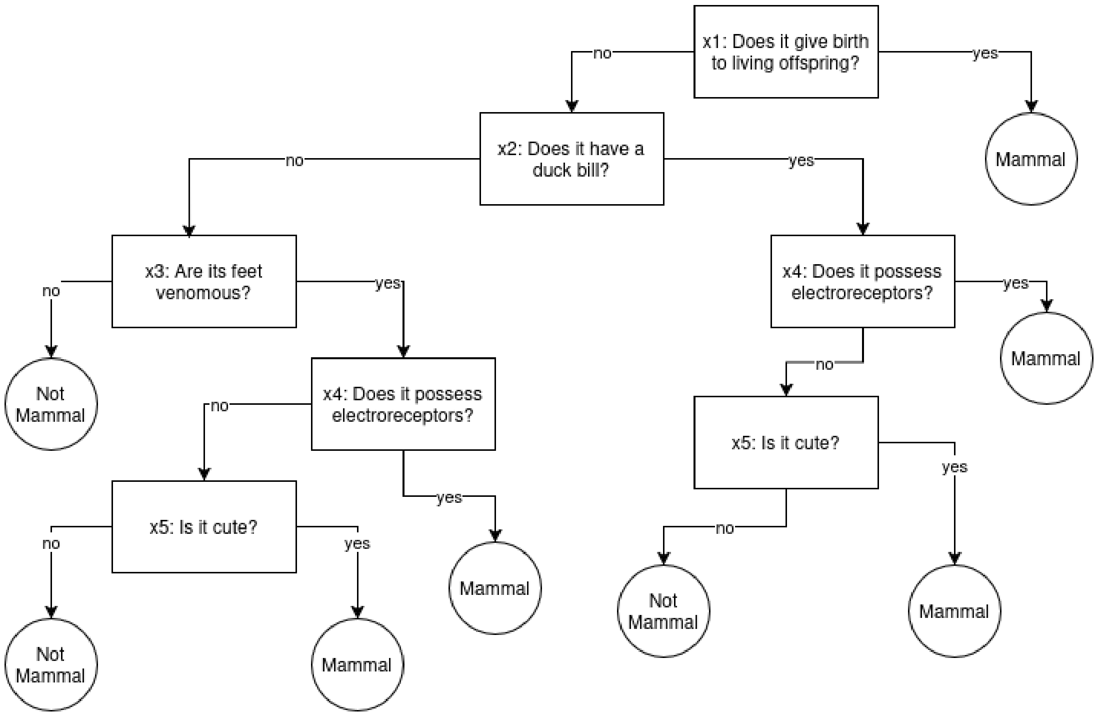


Figure 47: Adapted decision tree example[1]

[1]"Disproving XAI Myths with Formal Methods – Initial Results", Joao Marques-Silva

Abductive explanations

- $x = \{1, 0, 1, 0, 1\}$
- Why $f(x) = 1$ (mammal)?

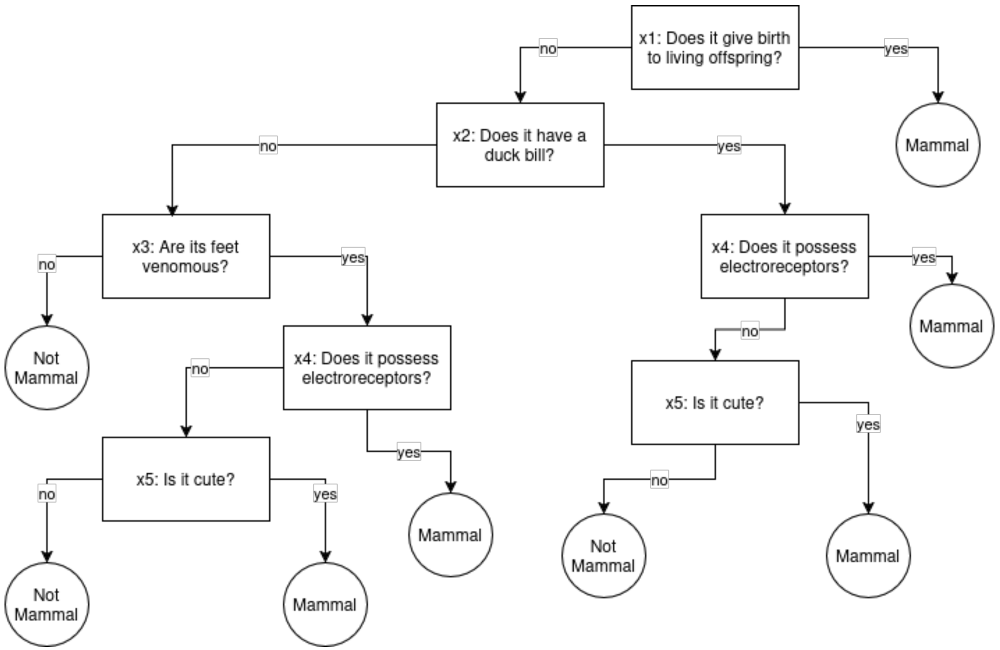


Figure 48: Adapted ecision tree example[1]

---

[1]"Disproving XAI Myths with Formal Methods – Initial Results", Joao Marques-Silva

Abductive explanations

- $x = \{1, 0, 1, 0, 1\}$
- Why $f(x) = 1$ (mammal)?
- An abductive explanation (AXP) is a subset-minimal of features $i \in \mathrm{AXP}$, such that:

$$\forall \overline{x} \in X. \left[\left(\bigwedge_{i \in \mathrm{AXP}} (\overline{x}_i = x_i) \right) \to f(\overline{x}) = c \right) \right]$$

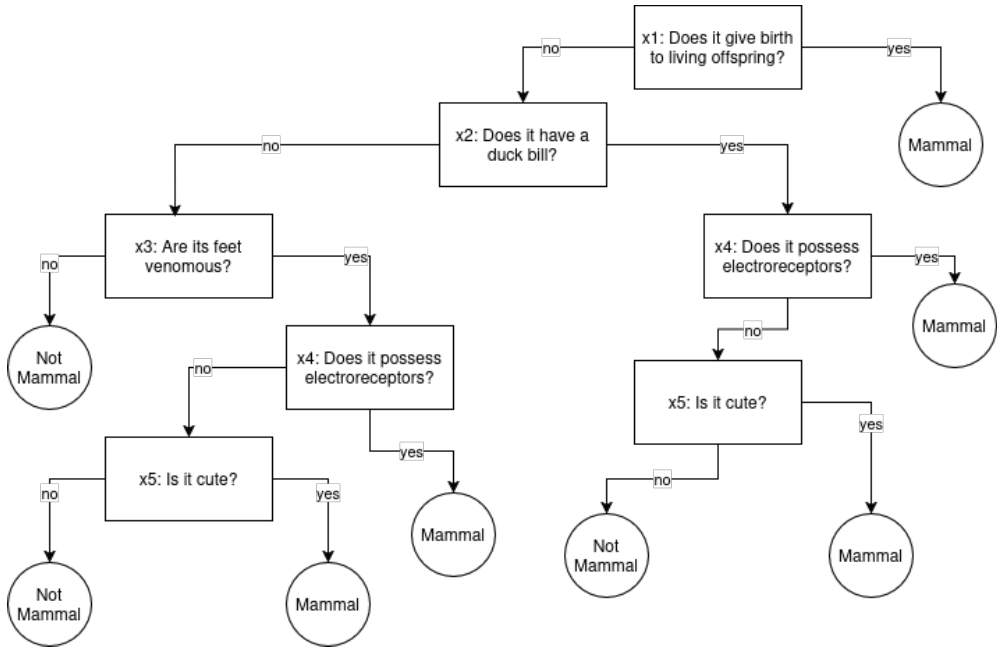


Figure 51: Adapted ecision tree example[1]

[1]"Disproving XAI Myths with Formal Methods – Initial Results", Joao Marques-Silva

Abductive explanations

- $x = \{1, 0, 1, 0, 1\}$
- Why $f(x) = 1$ (mammal)?
- An abductive explanation (AXP) is a subset-minimal of features $i \in \mathrm{AXP}$, such that:

$$\forall \overline{x} \in X. \left[\left(\bigwedge_{i \in \mathrm{AXP}} (\overline{x}_i = x_i) \right) \rightarrow f(\overline{x}) = c \right) \right]$$

- A weak abductive explanation (wAXP) is an AXP that is not subset minimal:

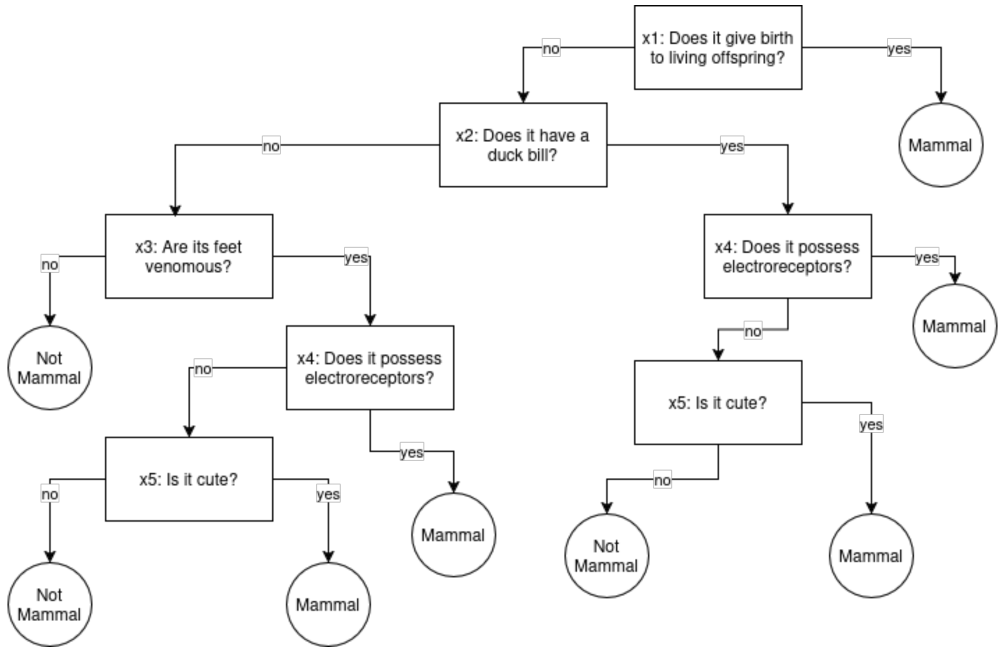


Figure 51: Adapted ecision tree example[1]

[1]"Disproving XAI Myths with Formal Methods – Initial Results", Joao Marques-Silva

Abductive explanations

- $x = \{1, 0, 1, 0, 1\}$
- Why $f(x) = 1$ (mammal)?
- An abductive explanation (AXP) is a subset-minimal of features $i \in \mathrm{AXP}$, such that:

$$\forall \overline{x} \in X. \left[\left(\bigwedge_{i \in \mathrm{AXP}} (\overline{x}_i = x_i)\right) \to f(\overline{x}) = c\right)\right]$$

- A weak abductive explanation (wAXP) is an AXP that is not subset minimal:
- Trivial wAXP: $\{x_1, x_2, x_3, x_4, x_5\}$

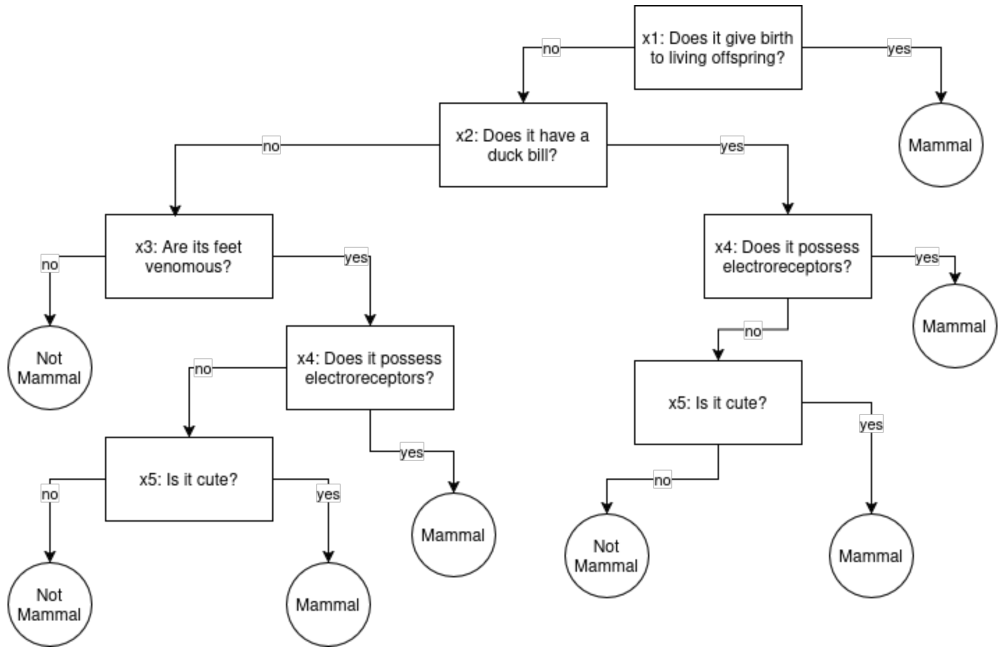


Figure 51: Adapted ecision tree example[1]

[1]"Disproving XAI Myths with Formal Methods – Initial Results", Joao Marques-Silva

Abductive explanations

- $x = \{1, 0, 1, 0, 1\}$
- Why $f(x) = 1$?
- An abductive explanation (AXP) is a subset-minimal of features $i \in \mathrm{AXP}$, such that:

$$\forall \overline{x} \in X. \left[\left(\bigwedge_{i \in \mathrm{AXP}} (\overline{x}_i = x_i)\right) \to f(\overline{x}) = c\right)]$$

- Subset minimal AXP: $\{x_1\}$

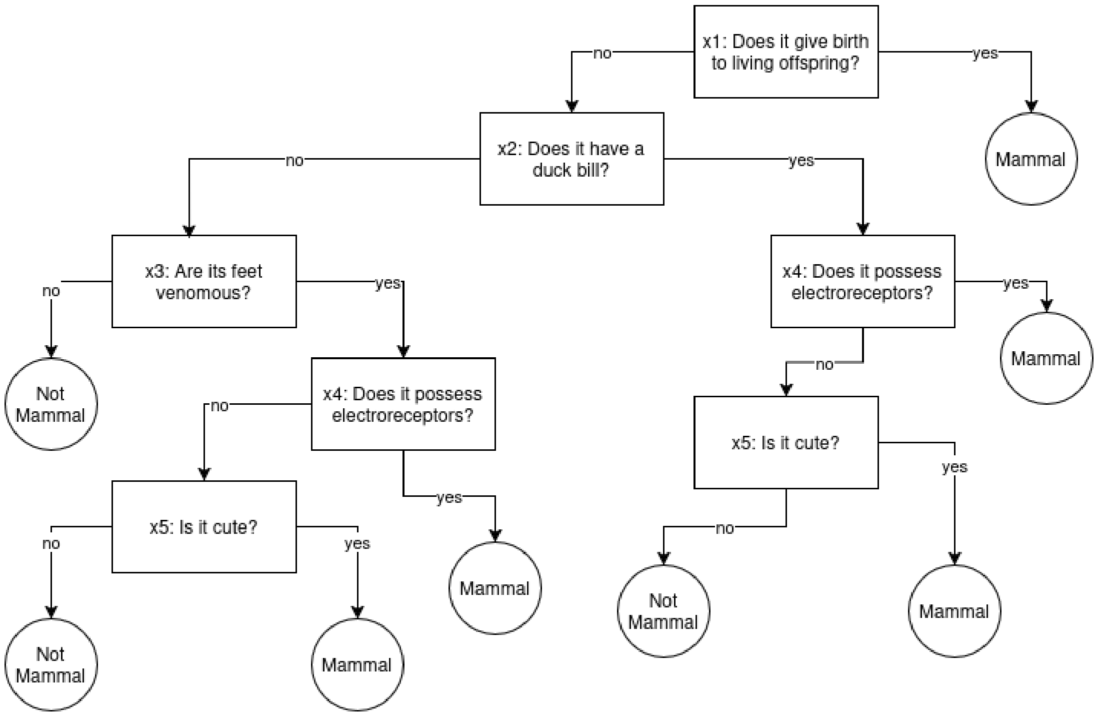


Figure 52: Adapted decision tree example[1]

[1]"Disproving XAI Myths with Formal Methods – Initial Results", Joao Marques-Silva

Algorithm to a formal explanation

Input: function f, datapoint $x = \{x_1, x_2, ..., x_n\}$

Output: Subset minimal abductive explanations

1. $\mathrm{wAXP} \leftarrow \{1, 2, ..., n\}$ (weak Abductive Explanation)
2. $c = f(x)$
3. for i in {1, 2, ..., n} do:
4. $\mathrm{wAXP'} \leftarrow \mathrm{wAXP} - \{i\}$
5. Check SAT of $\forall \overline{x} \in X. \left[\left(\bigwedge_{i \in \mathrm{wAXP'}} (\overline{x}_i = x_i) \right) \rightarrow f(\overline{x}) = c \right)]$
6. if SAT then $\mathrm{wAXP} \leftarrow \mathrm{wAXP'}$
7. $\mathrm{AXP} \leftarrow \mathrm{wAXP}$
8. Return AXP

Algorithm to a formal explanation

1. $\mathrm{wAXP} \leftarrow \{1, 2, ..., n\}$ (weak Abductive Explanation)
2. $c = f(x)$
3. for i in {1, 2, ..., n} do:
    4. $\mathrm{wAXP}' \leftarrow \mathrm{wAXP} - \{i\}$
    5. Check SAT of $\forall \overline{x} \in X. \left[\left(\bigwedge_{i \in \mathrm{wAXP}'} (\overline{x}_i = x_i)\right) \rightarrow f(\overline{x}) = c\right]$
    6. if SAT then $\mathrm{wAXP} \leftarrow \mathrm{wAXP}'$
7. $\mathrm{AXP} \leftarrow \mathrm{wAXP}$
8. Return AXP

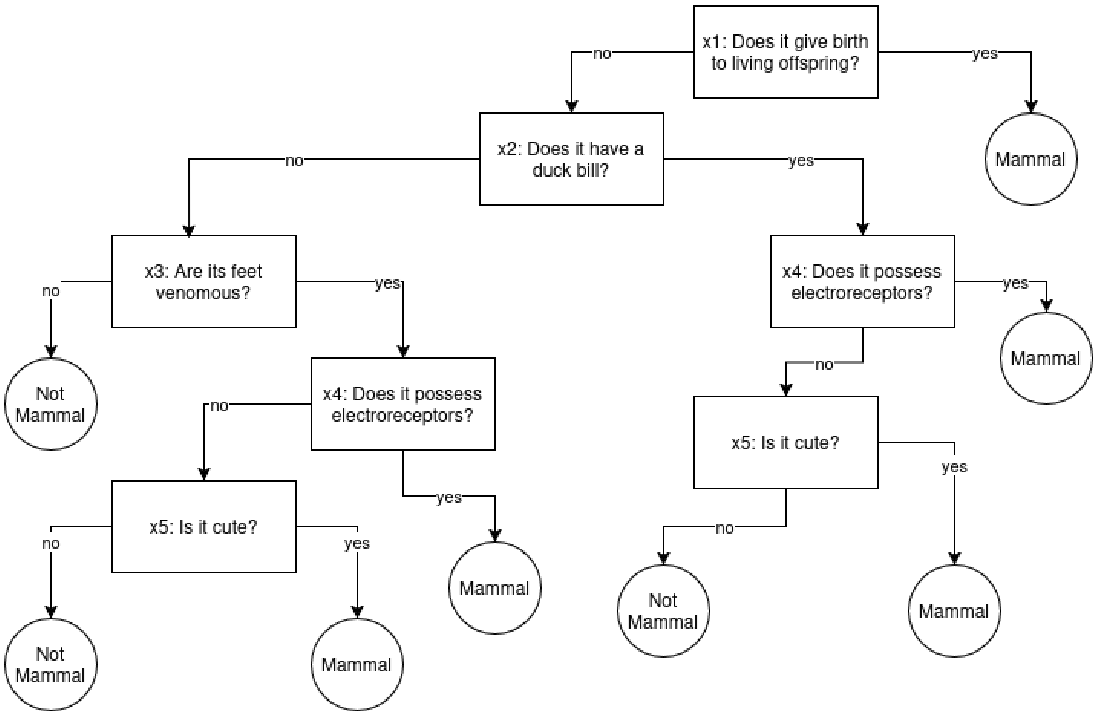


Figure 53: Adapted decision tree example[1]

[1]"Disproving XAI Myths with Formal Methods – Initial Results", Joao Marques-Silva

Contrastive explanation

- Similarly, we can ask “Why not “not mammal”?”

Contrastive explanation

- Similarly, we can ask “Why not “not mammal”?”
- A contrastive explanation (CXP) is a subset-minimal of features s.t.:

$$\exists \overline{x} \in X. \left[\left(\left(\bigwedge_{i \in (X - \text{CXP})} (\overline{x}_i = x_i) \right) \wedge f(\overline{x}) \neq c \right) \right]$$

Contrastive explanation

- Similarly, we can ask “Why not “not mammal”?”
- A contrastive explanation (CXP) is a subset-minimal of features s.t.:

$$\exists \overline{x} \in X. \left[\left(\left(\bigwedge_{i \in (X - \text{CXP})} (\overline{x}_i = x_i) \right) \wedge f(\overline{x}) \neq c \right) \right]$$

- Similarly, a weak CXP (wCXP) is a CXP that is not subset minimal.

Are CXPs and AXPs unique? (No)

- we should've seen in the algorithm that we got a different AXP than before
- it depends on the order in which you traverse the features
- AXPs: $\{1\}, \{3, 5\}$

[1]"Delivering Trustworthy AI through Formal XAI" by Joao Marques-Silva and Alexey Ignatiev

Are CXPs and AXPs unique? (No)

- we should've seen in the algorithm that we got a different AXP than before
- it depends on the order in which you traverse the features
- AXPs: $\{1\}, \{3, 5\}$

Duality between CXPs and AXPs: Minimal hitting sets[1]

- Interesting property: If you were to find all AXPs, it would allow you to compute a CXP aswell!
- CXPs: $\{1, 3\}, \{1, 5\}$

[1] "Delivering Trustworthy AI through Formal XAI" by Joao Marques-Silva and Alexey Ignatiev

Are CXPs and AXPs unique? (No)

- we should've seen in the algorithm that we got a different AXP than before
- it depends on the order in which you traverse the features
- AXPs: $\{1\}, \{3, 5\}$

Duality between CXPs and AXPs: Minimal hitting sets[1]

- Interesting property: If you were to find all AXPs, it would allow you to compute a CXP aswell!
- CXPs: $\{1, 3\}, \{1, 5\}$

Open Problem: Multiple explanations for the same decision

- Given multiple formal explanations, which one to choose?

[1] "Delivering Trustworthy AI through Formal XAI" by Joao Marques-Silva and Alexey Ignatiev

New paradigm: Local formal explanations

Reminder:

Local robustness (Katz et al. 2017)

Let a classifier $f : \mathcal{X} \mapsto \mathcal{Y}$. Given $x \in \mathcal{X}$ and $\varepsilon \in \mathbb{R} << 1$ the problem of *local robustness* is to prove that $\forall x^{\{'\}}.\ \|x - x^{\{'\}}\|_p < \varepsilon \to f(x) = f\left(x^{\{'\}}\right)$

Algorithm to a formal explanation (VeriX)

Input: function f, datapoint $x = \{x_1, x_2, ..., x_n\}$, *perturbation level* ε

Output: Subset minimal local abductive explanations

1. $\mathrm{wAXP} \leftarrow \{1, 2, ..., n\}$ (weak Abductive Explanation)
2. $c = f(x)$
3. for i in {1, 2, ..., n} do:
    4. $\mathrm{wAXP'} \leftarrow \mathrm{wAXP} - \{i\}$
    5. Check SAT of:

    $$\forall \overline{x} \in X. \left[\left(\bigwedge_{i \in \mathrm{wAXP'}} (\overline{x}_i = x_i) \wedge \left(\bigwedge_{\boldsymbol{j} \notin \mathbf{wAXP'}} (\|\overline{\boldsymbol{x}}_{\boldsymbol{j}} - \boldsymbol{x}_{\boldsymbol{j}}\|_{\boldsymbol{p}} \leq \boldsymbol{\varepsilon}) \right) \rightarrow f(\overline{x}) = c \right) \right]$$

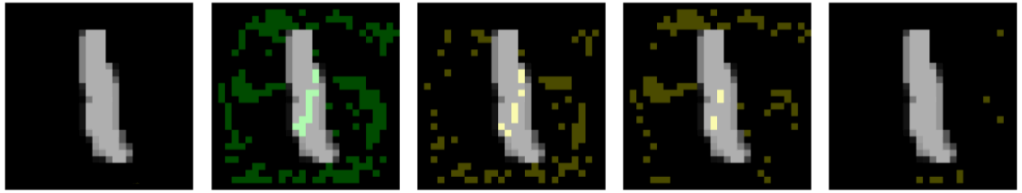
    6. if SAT then $\mathrm{wAXP} \leftarrow \mathrm{wAXP'}$
7. $\mathrm{AXP} \leftarrow \mathrm{wAXP}$
8. Return AXP

Applicable domains

- We've seen FXAI for decision trees and tabular data
- They can also work on more complex scenarios: neural networks, computer vision
- While some problems are quick to solve (instantanous), complex problems take a long time:

Example explanations

- Explanations computed for the decision of a convolutional neural network



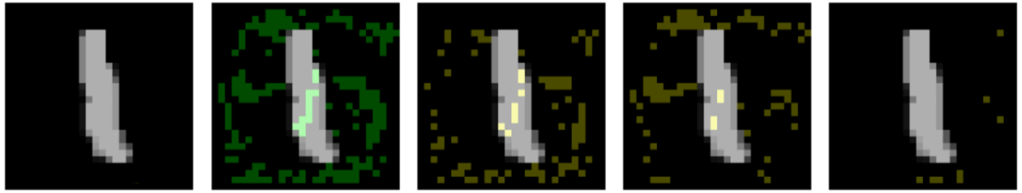
(b) Handwritten digit "1", not "8", not "5", not "2"

Figure 57: Explanation from VeriX[1]. Image resolution 1x28x28

[1] "VERIX: Towards Verified Explainability of Deep Neural Networks", by M. Wu, H. Wu and C. Barrett

Example explanations

- Explanations computed for the decision of a convolutional neural network
- Question: What is your take on this explanation?



(b) Handwritten digit "1", not "8", not "5", not "2"

Figure 60: Explanation from VeriX[1]. Image resolution 1x28x28

[1] "VERIX: Towards Verified Explainability of Deep Neural Networks", by M. Wu, H. Wu and C. Barrett

Example explanations

- Explanations computed for the decision of a convolutional neural network
- Question: What is your take on this explanation?

Figure 61: Example of original image (left) and of the explanation (right) for the current prediction (give way sign). Image resolution: 3x32x32

Example explanations

- Explanations can be computed for larget datapoints, on larger networks
- Question: What is your take on these explanations?

Figure 62: Example of original image (top-left) and of three different explanations for the same decision (“Watch”) on Resnet18 (same weights). Image resolution: 3x64x64

Practicality

Dataset / Method	MNIST	GTSRB	GTSRB, simplified	TinyImageNet, simplified
Satisfiability Modulo Theories	820s	2400s	32s	-
Abstract Interpretation	220s	1600s	8s	4050s
Interval Bound Propagation	40s	140s	1s	-

Limitations of formal XAI

From why to use it, now we go back to what are limitations for it:

- tabular data (or smaller in size) vs bigger data (eg visual)
- Scalability, memory usage, viability of explanations
- NLP tasks & transformers, approximating various activation functions
  - (eg Softmax is an unsolved problem that prevents us from verifying transformers)
  - (eg the discreteness of words is an unsolved problem that prevents us from verifying NLP models)
- + your opinions of the formal explanations you have seen

Open ended directions

- Disentangle the multiplicity of explanations
- Scalability
- Apply FXAI to multicriterion decision making (analysis of conflicting criteria)

Key takeaways from lecture

- Explainability is important in Artificial Intelligence
- Explanations take into account both the system explained and the user seeking the explanation
- No XAI method is flawless
- Formal methods address limitations in terms of model faithfulness and explanation compactness

Feedback

Thank you for listening to me!