Deep Explanation

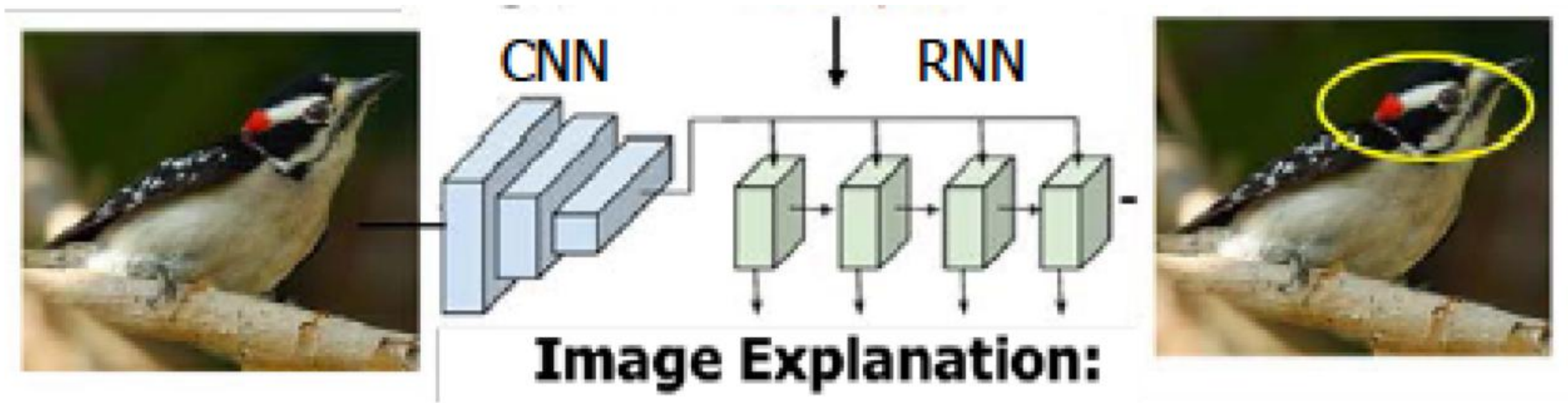
Sargur N. Srihari
srihari@buffalo.edu

Topics in Deep Explanation

1. Embedding Deep Nets in Visual Explanation
2. Visual Saliency
3. Bayesian Deep Learning
4. Criticisms of Ante-hoc AI

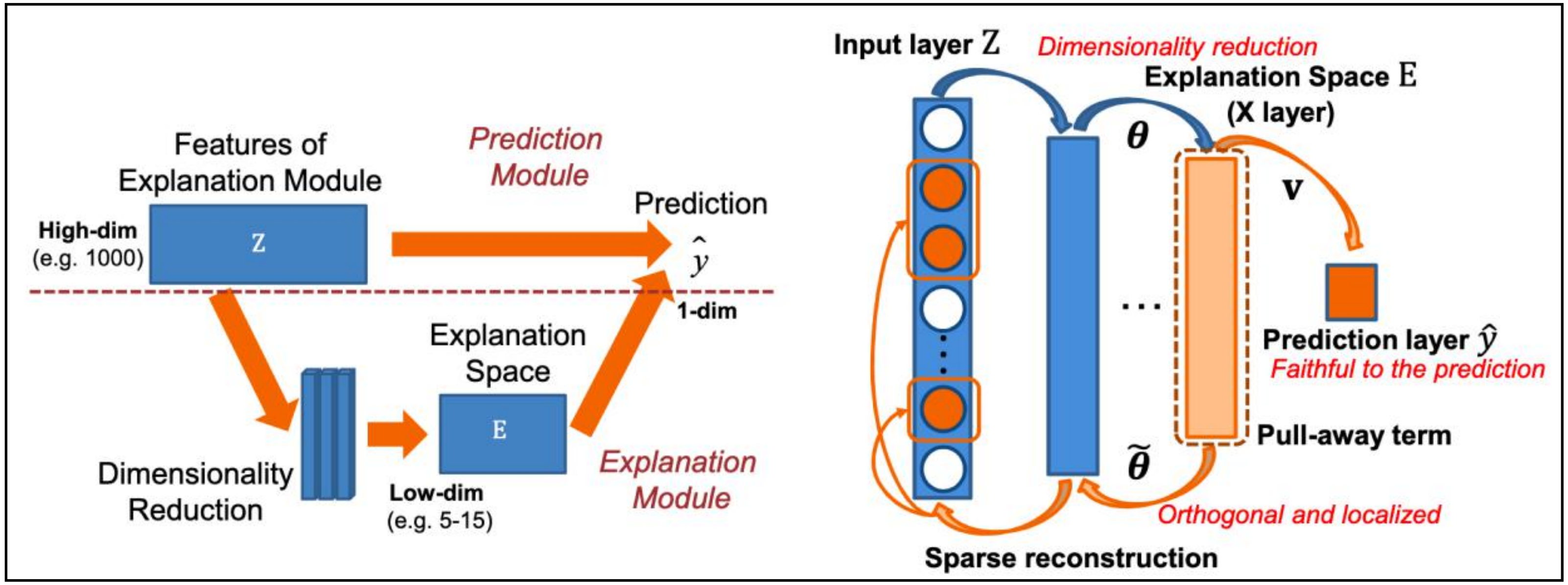
Deep Explanation

Downy Woodpecker definition:
This bird has a white breast, black wings and a red spot on its head



This is a Downy Woodpecker because it is a black and white bird with a red spot in its crown

Embedding Deep Networks into Visual Explanations

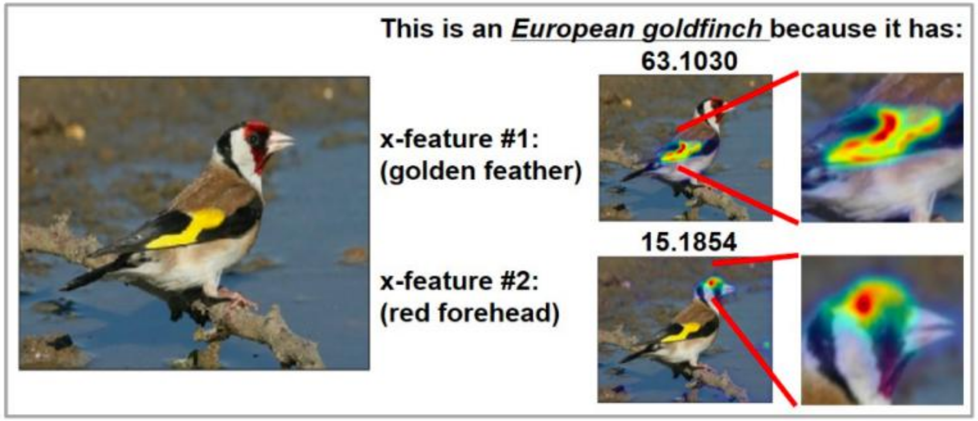


Explanation module is a dimensionality reduction mechanism so that the original deep learning predictionˆy can be reproduced from this low-dimensional space.
It can be attached to any layer in the prediction deep network (DNN)
The DNN output can be faithfully recovered from this low-dimensional explanation space

Sparse Reconstruction Autoencoder is used a explanation module

Embedding Deep Networks into Visual Explanations

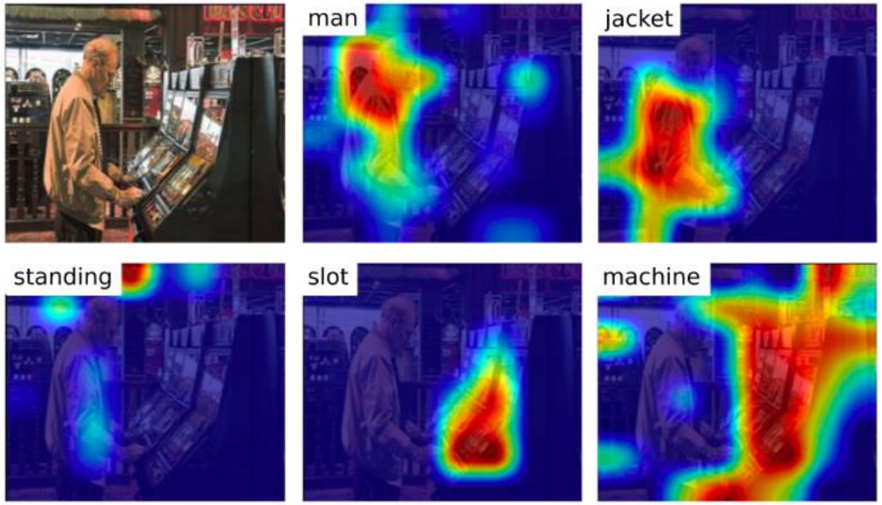
People prefer explanations of the form ‘A” is something because of B,C and D
This is a bird because it has feathers, wings and a beak
It is concise– here are not a hundred reasons
It relies on B,C,D which are also high level concepts



Approach generates visualizations for humans to deduce those features
Without requiring textual annotation

Explaining Images

- Deep image captioning systems
  - – learn to translate visual input into language
    - potential map between visual concepts and words
  - – Despite good captioning performance, they are hard to understand “black boxes.”
- Solution: Caption guided visual saliency
  - – Top-down neural saliency map

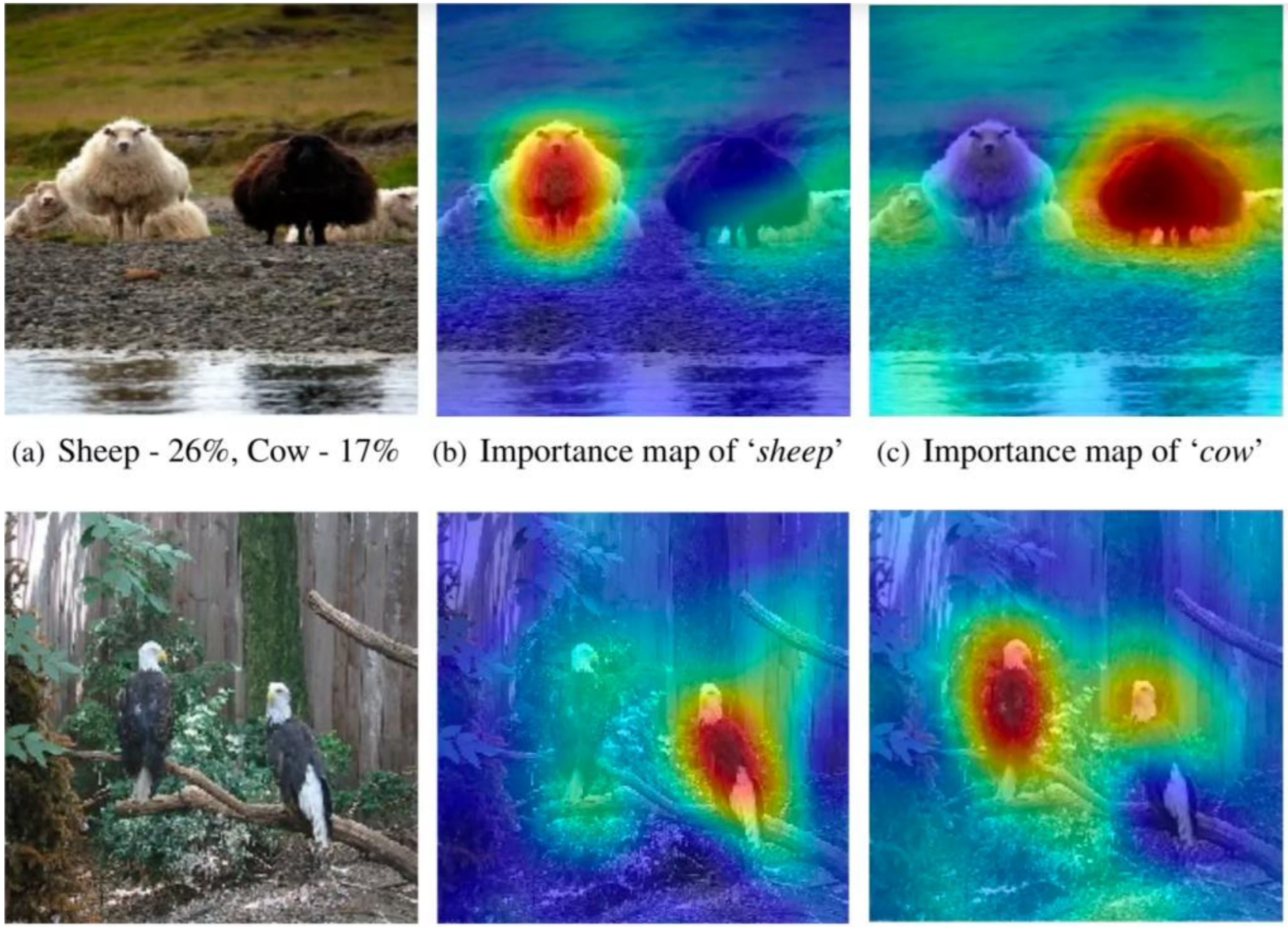


– Input:

– A man in a jacket is standing at the slot machine

Saliency Maps Produced by RISE

https://bdtechtalks.com/2018/10/15/kate-saenko-explainable-ai-deep-learning-rise/?fbclid=IwAR1RrH-BrLqRnXMiqcvAv3MtXQ6AcBAdufvWPROncF2rzPslOfoU5SD-2bM

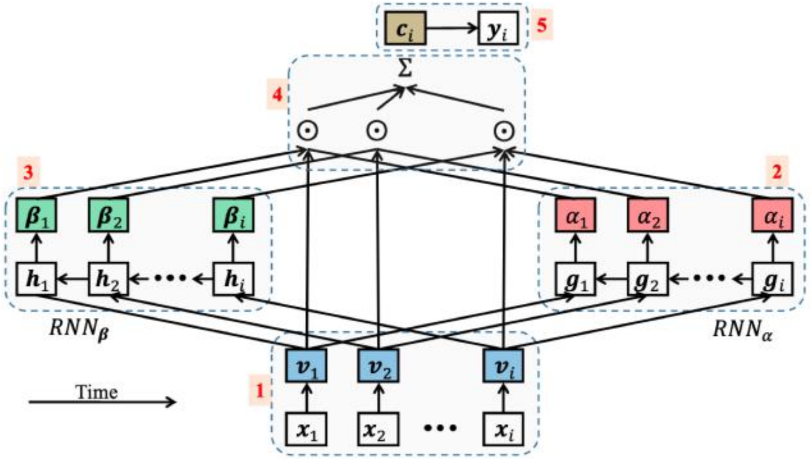


(a) Sheep - 26%, Cow - 17% (b) Importance map of '*sheep*' (c) Importance map of '*cow*'

(d) Bird - 100%, Person - 39% (e) Importance map of '*bird*' (f) Importance map of '*person*'

An ante-hoc system: RETAIN

- Reverse time Attention Model
  - Mimics physician: using EHR in reverse time order
  - Calculates contribution of the variables (medical codes) to diagnostic prediction using RNNs



Source: Choi, etal, NIPS 2016

Attention in Machine Translation

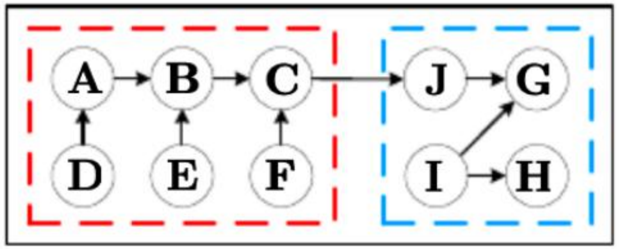
Given sentence of length S in the source, generate $h_1,\ldots,h_S$, to represent input words. To find j^{th} target word, generate attention $\boldsymbol{\alpha}_i$ for $i=1,\ldots,S$ for each word in source sentence.

Compute context $c_j=\Sigma_i\ \boldsymbol{\alpha}_i^j\ h_i$ and use it to predict j^{th} target word i

Attention allows model to focus on specific words in given sentence when generating each word in the target

Ante-hoc: Bayesian Deep Learning (BDL)

- Tightly integrating deep learning with a PGM
- Medical diagnosis example
  - After seeing visible symptoms (images) infer etiology (causation) from all symptoms
  - Reasoning is beyond deep learning models
  - PGMs are poor at perceptual tasks (but readily generate explanations)

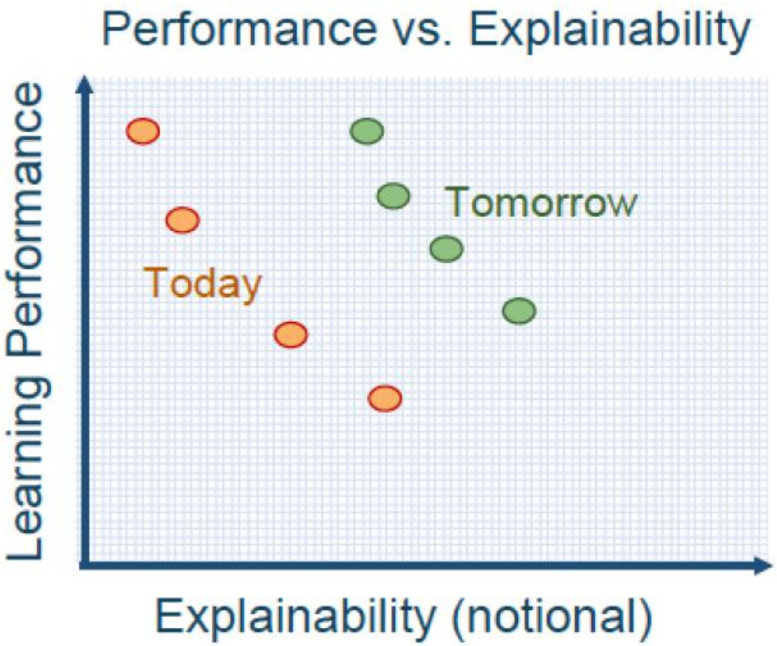


Red rectangle is perception component
Blue rectangle is task-specific component
J is the hinge variable

Source: Wang, Yeung IEEE-KDE

Criticisms of Ante-hoc XAI

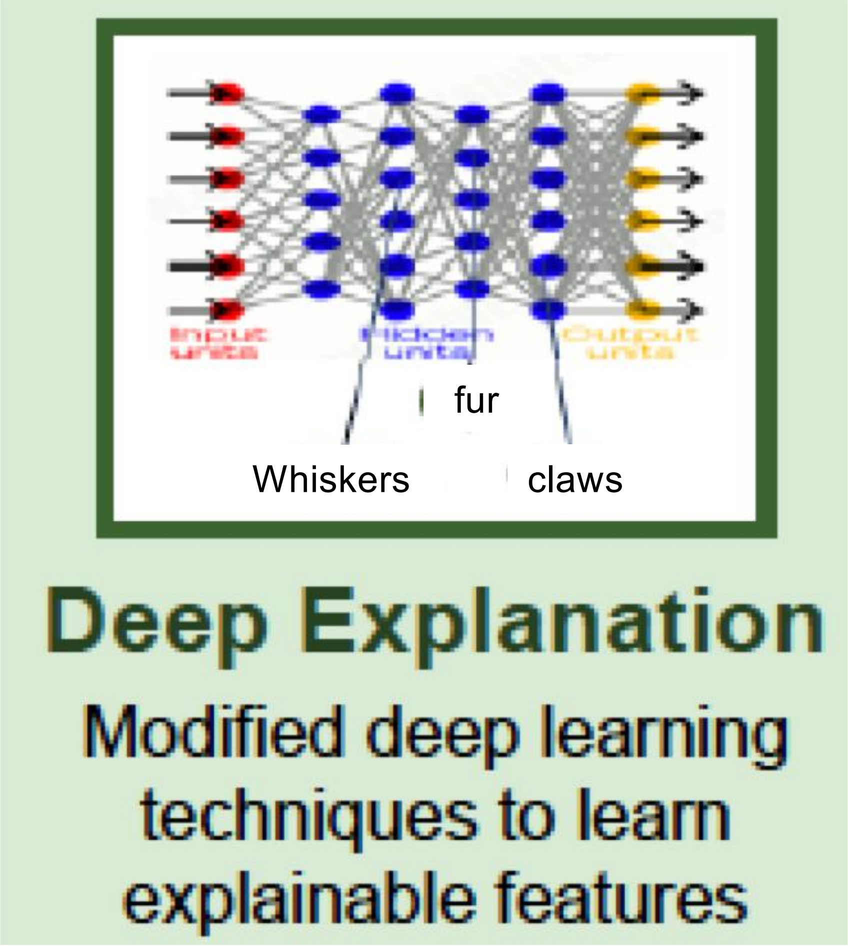
1. A complicated blackbox does not necessarily have the best performance
   - Deep neural nets and logistic regression have same performance with principled feature selection



Criticism of ante-hoc XAI

2. XAI unfaithful to original model

- If XAI agrees with model 90%, it is wrong 10%
  - Thus the original model is not trustable



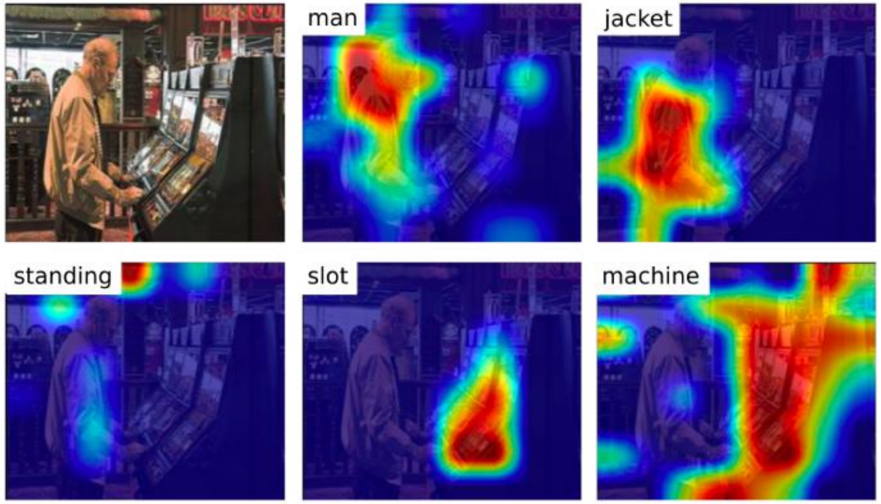
Deep Explanation

Modified deep learning techniques to learn explainable features

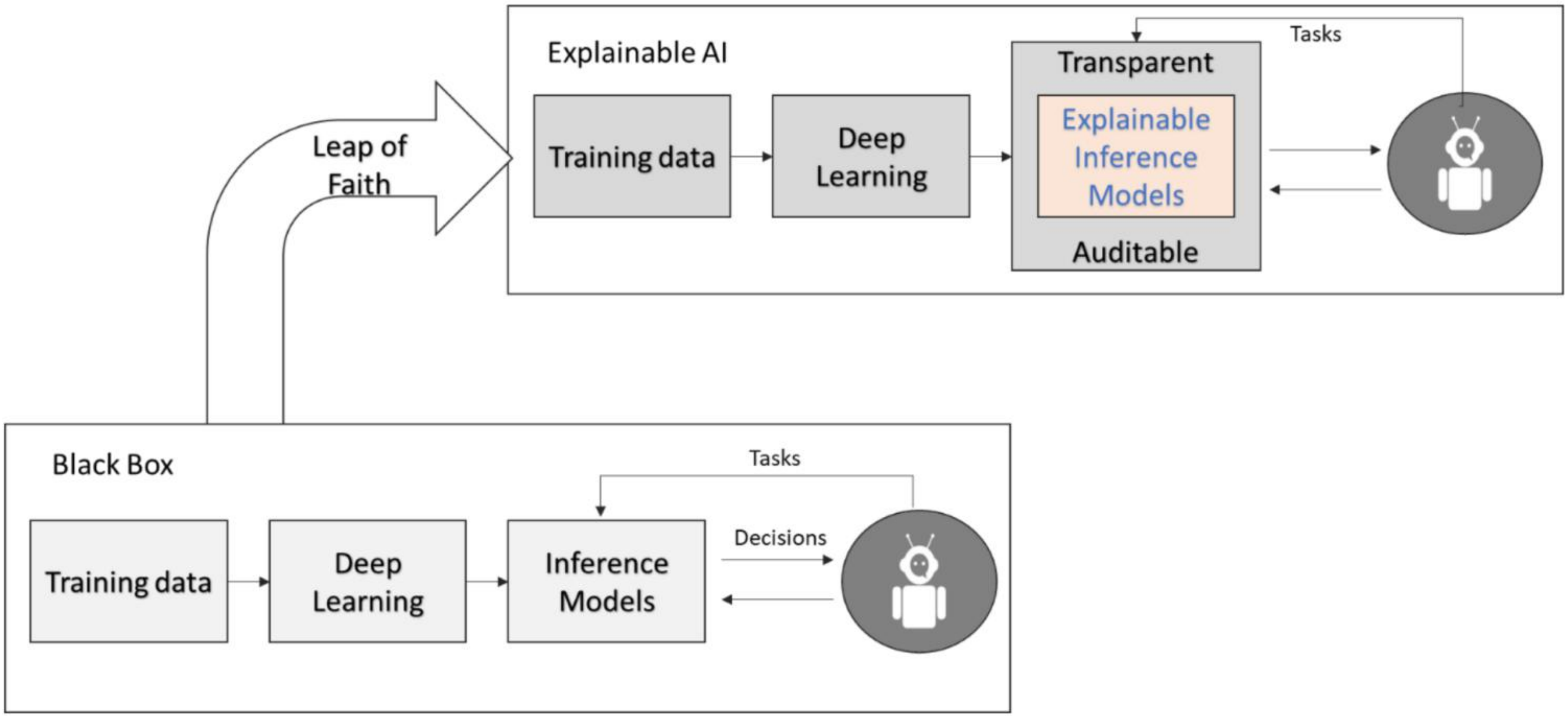
Issues in XAI

3. Explanations may be incomplete

- Saliency may does not say how it is being used



Ante-hoc XAI is a leap of faith



Source: https://www.hcltech.com/blogs/explainable-artificial-intelligence-inflection-point-ai-journey

Explainability: human introspection

- If an algorithm could self-explain it would be like asking a human to introspect
- They would simply make up a story
- Emotive vs Non-emotive content
  - With emotive content
    - Q: Why did you throw the plate?
    - A: Because of childhood trauma (from *limbic* system below consciousness)
  - Non-emotive question
    - Q: Why did you classify as a dog
    - A: It has 4 legs, tail, cylindrical body, all arranged spatially