How can we make Deep Learning more transparent?

John Hios

1. Intro
2. What is Deep Learning?
3. Ethical Challenges/Ambiguities
4. Approaches to Solve these Issues

Intro

Problem Statement:

Deep neural networks are notoriously opaque to human inspection. It's difficult for users to understand how they arrived at their output.

As deep learning algorithms are being used for applications from facial recognition to healthcare diagnoses to autonomous vehicles, it is increasingly more important for their inputs and outputs to be **transparent**.

Questions Posed:

- Are there situations where deep learning should never be used?
- Are there additional safeguards required?

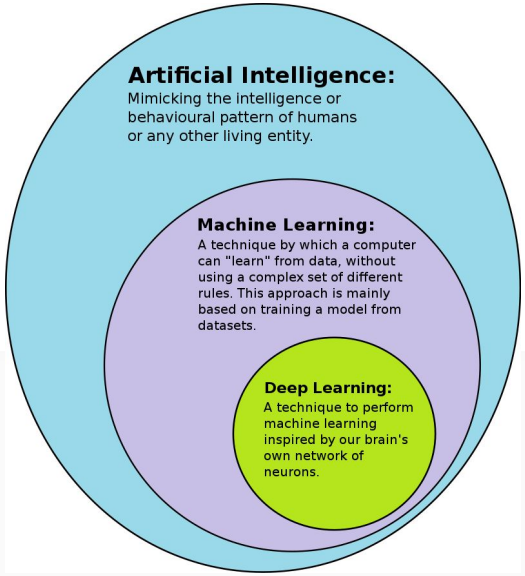
What is Deep Learning?

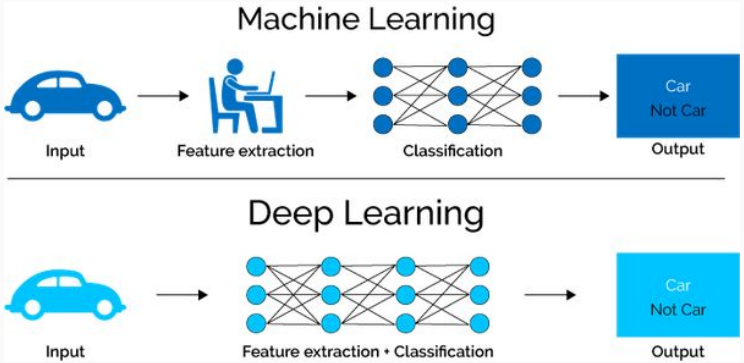
Deep Learning is a subset of machine learning where artificial neural networks, algorithms inspired by the human brain, learn from large amounts of data.

Deep Learning allows machines to solve complex problems even when using a data set that is very diverse, unstructured and inter-connected.

The more Deep Learning algorithms learn, the better they perform (ideal for Big Data).

Applications: Image/facial recognition, Drug discovery & toxicology, Medical image analysis, Autonomous vehicles, Military

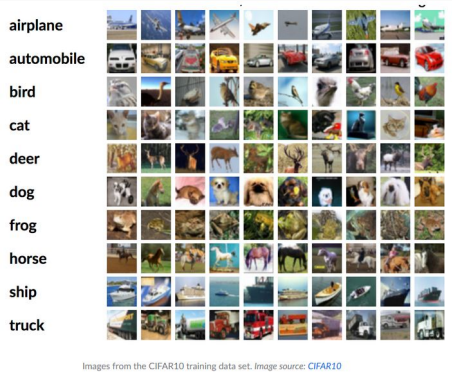




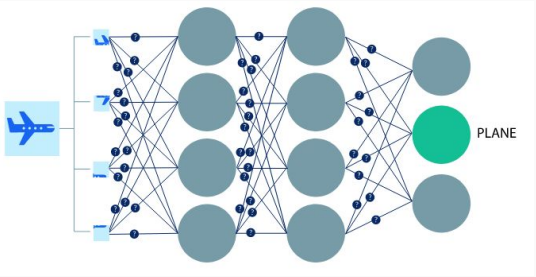
Ethical Challenges/Ambiguities - Lack of Transparency

Humans are left out of the loop

Difficult to understand and explain these models



Images from the CIFAR10 training data set. *Image source:* CIFAR10

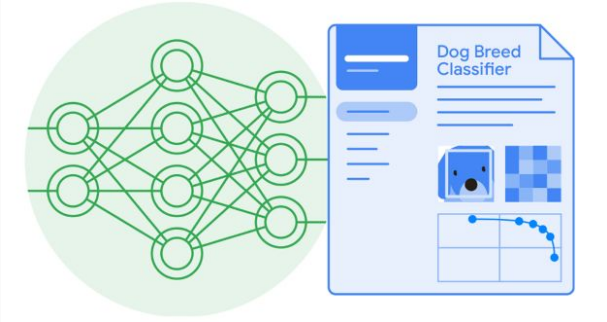


Problems associated with lack of transparency

1. Unexplainable algorithms
2. Lack of visibility into training data sets
3. Lack of visibility into methods of data selection
4. Limited understanding of the bias in training data sets
5. Limited visibility into model versioning
6. Model trust (why was the model built in the first place? Are we using it the way it was designed for? …..)

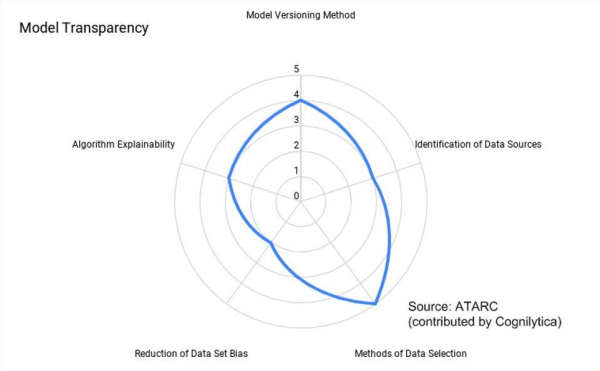
(Some) Approaches to Solve Transparency Issues

Google Cloud Models Cards



Primarily focused on helping the model builder build better models, but not necessarily focused on third party model consumers.

ATARC Model Transparency Assessment Chart



These assessments are supposed to be produced from the model builders - no bias(?)

Thanks!