Interpretability and Explainability as Necessary Pieces for Machine Ethics

Implementing Machine Ethics Workshop

Taghi Aliyev, Marco Manca, Mario Falchi, Alberto Di Meglio

02/07/2019
About Me

Brief Introduction

Taghi Aliyev, Implementing Machine Ethics
Ethics and Sustainability

Limitations, challenges, problems

• Limited negotiation powers in decision-making
  • With Deep Learning and other recent ML-based systems

• Not all the outputs understood or explained

• Sustainability issues:
  • Need for rules and law updates constantly

• Ethical challenges:
  • Biased systems
  • Unavailability of explanations or explicit correlations
Reasonable Inference
Fences and regulations

Taken from Pasquale Cirillo's "Of Risk, Fences and unavoidable falls"
Explainable AI

• Ongoing preference towards non-"black box" models

• A paradox with recent advances:
  • Better models are available, however preferred to simpler models

• Explainability of the black box models
  • Open topic
  • Next necessary step in sustainability of Deep Learning models
  • Error correction
Explainable AI

Tendencies towards non Deep Learning approaches

A White-Box Machine Learning Approach for Revealing Antibiotic Mechanisms of Action

Algorithms and Justice: Scrapping the ‘Black Box’

By Cynthia Rudin | January 26, 2018

Taghi Aliyev, Implementing Machine Ethics
Explainable AI

Required characteristics

• Negotiate the inference

• Provide useful new insight from complex modelling techniques

• Meaningful human control over the systems

• Similar with scientific findings/hypothesis:
  • If a scientist produces a new theory or finding, they need to prove it and explain it
  • Same should be upheld for ML systems
Explainable AI

Twins UK Study together with King’s College London

• A step towards interpretability and meaningful human control

• Detection of facial features in twins
  • Initially heritability analysis
  • Next clinical traits and disease symptoms

• Adaptive pipeline for deconvolution
  • Based off the work from M. Zeiler (2014)
Explainable AI

Adaptive approach for Twins UK Study

• Optimize the approach proposed by Zeiler

• Include clinical information of interest in selection of interesting neurons

• Refined approach
Explainable AI

Promising results
Limitations and expandability

• Not general enough
  • Currently, working on CNNs

• Expandability:
  • Better approximation approaches -- Exploring search space
  • Testing and adjusting for more complex networks
  • Different when comes to RNNs and attention networks

• Pro:
  • Working on pre-trained network on a general data set
Outlook and closing remarks

• Importance of Causal Reasoning
  • Cornerstone idea for the explainable AI

• Cultural difference within AI
  • Curve Fitting/Association learning vs Casual Reasoning

• TwinsUK study an important step for explainability in Medical Research
  • Practical study and example
Acknowledgement

Meet the team!

Dr. Mario Falchi, KCL

MD, Marco Manca, SCImPULSE

Dr. Alberto Di Meglio, CERN openlab