Explanations from Gods of Olympus to black-box models: a layman perspective

Pietro Baroni

DII - Dip. di Ingegneria dell’Informazione

University of Brescia (Italy)

explAIn Technical Workshop

July 8, 2021

The jungle of XAI

- The pressing needs and huge interest concerning XAI need not to be commented
- Many ways and approaches to produce explanations …
- … in spite of the lack of a "clear" notion of explanation
- *it is fair to say that most work in explainable artificial intelligence uses only the researchers' intuition of what constitutes a ‘good’ explanation* [Miller, AIJ, 2019]
- Many terms used with non uniform meaning like interpretability, fidelity, faithfulness, …
- User experiments have possibly "the final word"

In search of roots

- Can we attack the question "from the roots"?

- Explanations have evolved in human history:
  can a historical perspective give any useful indication or at least any hint?

- A layman (in all possible respects) perspective shared for discussion

- Let us consider a phenomenon which strikingly powerful and marvelous but also frightening

Explaining thunderbolts

- In ancient Greece there was a simple explanation for thunderbolts

- Jupiter (aka Zeus) is throwing them!

Jupiter better than nothing?

- Did ancient Greeks really believe in this explanation?
- In any case, they generated and somehow used this imaginative explanation!
- Why did they?
- Some layman hypotheses

Horror vacui: no empty explanations

- It seems that we as humans are uncomfortable with (or are afraid of or have other uneasy feelings about) things that just happen
- We feel the need of making sense of things using entities and categories belonging to our "cognitive sphere"
- An explanation in terms of such entities is better than nothing ("unbelievable" or "unverifiable" as it may be)
- The idea of a human-like subject doing a common action (throwing things) is more acceptable than no idea at all

Illusion of (potential) control: actionable explanations

- Once there is an explanation one can think that some action is possible in order to have an influence on the considered phenomenon

- If thunderbolts just happen: nothing to do

- If Jupiter is throwing them: one can offer sacrifices to him to get its benevolence

Gods as "universal" explanations

- The gods-based explanation scheme is versatile and multipurpose
- Gods' actions were used to explain both natural phenomena (winds, earthquakes, epidemics) and otherwise unexplainable individual behaviors (e.g. Venus governing love affairs "from outside")
- A fairly "robust" and "satisfactory" explanation scheme
- Probably it would have passed any user test

Psychological motivations of ancient explanations

- Reasons like "horror vacui" and illusion of control look more related to psychological needs than to a "rational" attitude

- Ill-founded as they are, these explanations may have been useful for people to come to terms with an otherwise untolerable reality

- Psychologically useful explanations are practically useful: if they supported people's well-being at some point, they supported progress in the long term

Explaining thunderbolts today (from Britannica.com)

As early as 1752, Benjamin Franklin discovered that lightning was caused by powerful electrical discharges in the clouds. Thunderstorms are caused by small electrically-charged particles. It's quite simple really. Water in clouds moves upwards. As it does so, it cools and freezes. As these ice particles fall back down they make contact with water droplets lower down in the cloud and this results in a charge separation. Within the cloud, two poles form, each with a different electrical charge.

On the ground, too, there are differences in electrical charges. Nature, though, is always striving to balance out these differences in electrical charges. That means that charged particles will always flow in the direction where there are less particles of the same charge. The result is a lightning bolt. At first, there is a bolt which is invisible to our eyes. At the same time, an excess of positively charged particles builds up on the ground, seen here in green. When the invisible lightning bolt gets close enough to the ground, there is a powerful discharge of energy. So powerful, in fact, that it results in an electrical arc. This is the lightning bolt that we see.

While this is happening, the surrounding air is heated to extreme temperatures. It expands and explodes, producing a loud crack. This is the thunder that we hear.

Lightning bolts come in many different colors. The color depends on atmospheric humidity, temperature and levels of air pollution. Depending on the circumstances, it might be red, blue or yellow. Lightning bolts are the hottest things on earth. They don't just heat the air to extreme temperatures, they also transport massive amounts of energy. They carry energy measuring several hundred billion watts. This is what makes lightning strikes so dangerous.

Jupiter vs. Britannica.com

- The explanation by Britannica is accessible only with some background knowledge in physics
- The explanation by Britannica is much more complex and articulated
- It is anyway appreciable by most people because the average "cognitive sphere" includes more scientific notions than in the past
- No illusion of actionability is left with Britannica (maybe some hope for the future is left open)

Jupiter vs. Britannica.com

- The Britannica explanation is still a high level and relatively rough description of the phenomena underlying thundebolts
- Most of the readers are happy with such a description and have no deeper knowledge on the matter (e.g. equations describing the behavior of electrical charges or the movements of water drops)
- We are confident that there are other more competent people able to deal with more detailed models and explanations if needed

Jupiter vs. Britannica.com

- Britannica helps making sense of thunderbolts with a greater sense of realism but …
- … even Britannica cannot provide a precise explanation of why a specific thunderbolt happened at a given moment and stroke exactly that bell tower
- It is acknowledged that the detailed explanation of single instances of thundebolts is beyond the current state of knowledge …
- … an maybe beyond some theoretically or practically unachievable complexity limits

What about Black-Box models?

- They too are strikingly powerful and marvelous but also frightening
- Real hype about explaining them (in some sense)
- Motivated by their increasing pervasiveness (if not intrusiveness) in many aspects of our life
- Intriguing challenge because Black-Box (BB) models are extremely complex and often untamable even by experts
- Right to explanation (whatever this means) sanctioned by law

Back to Jupiter ?

- Are the motivations for explaining BB-models psychological rather than rational?
- Is there an "horror vacui" about things that "just work"? (provided that they work)
- Need of making sense of things using entities and categories belonging to our "cognitive sphere", independently of their actual explanation value in "objective" terms
- Linear approximations and feature attribution (weighing) methods are assumed to belong to the cognitive sphere of most users but …

Back to Jupiter ?

- Increasing concerns about the inherent limitations of feature attribution methods
- They may fail to satisfy very basic requirements …
- … but they are better than nothing, aren't they?

- Model-agnostic explanation methods are universal, versatile and multipurpose … as Gods of Olympus were
- Not as imaginative and certainly more mathematically founded but …
- … which actual user needs are they addressing and under which hypotheses?

The illusion of actionability?

- Another major trend concerns counterfactual explanations
- Extremely intuitive, they support the considerations of alternative scenarios and the identification of decisive factors
- Natural interpretation in terms of what should I change (the rest being the same) to get a different outcome from the black-box but …
- It has been observed that at least in some cases this interpretation is highly misleading:
  - » changes may have side effects
  - » some combinations of conditions are unattainable in the real world
  - » the time required to act may per se prevent the rest being the same

Beyond Britannica?

- The style and depth of the explanation about thunderbolts provided by Britannica could be easily reproduced in a high-level explanation of the inner working and training of a black-box model …
- … but this is not enough
- Why are we much more demanding about black-box models than about other entities or phenomena?
- Why do we put such high (and possibly unreasonable) expectations on BB-models only?
- Focus on providing a sufficiently detailed account of every instance instead of understanding the inner behavior at a general level

Reasons to be demanding

- Some applications of the black-box models involve severe consequences on human life depending on their outcomes: they need to be explainable
- But it is known that the human experts we generally trust may use vast amounts of tacit knowledge and be unable to provide a "real" account of the motivations underlying their evaluations or decisions
- Other kinds of technologies are used in life-critical contexts and we commonly rely on people developing and testing them, rather than asking for explanations
- So, what's the difference?

Reasons to be demanding: irrational fears

- There might be an excessive attention on any system somehow related to Artificial Intelligence
- This term raises enthousiasm but also some (more or less justified) fear and suspicion
- AI artifacts seem to be treated differently from other technical artifacts maybe because they give the impression of replacing humans in a more radical way
- But other systems replacing humans exist, maybe we are just more accustomed to them

Reasons to be demanding: rational differences

- Some black-box models are perceived as differing substantially from other human-replacing techniques

- Not even experts can really understand them in all details…
- … but this can apply to many complex artifacts and techniques (from airplanes to vaccines)

- While generally accurate, they can make gross mistakes that no human would make
- … this seems to me the most crucial and most justified point
- … and the main one to work on

Some layman suggestions: education vs. explanation

- Reducing the behavior of very complex systems to explanations which are accessible to non experts and "close enough" to the working of the system is almost hopeless
- This means providing to final users explanations of "serious" black-box systems is almost hopeless
- In the long term, Britannica-style explanations could be achieved if the "cognitive sphere" of the users will include some basic notions of the principles ruling the operation of the black-box systems (then grey-box?)
- Education/persuasion in addition to explanation (consider vaccination campaigns)

Some layman suggestions: designers and experts in the loop

- Trust in designers and experts is a key point for trust in systems …
- Provided that the designers and BB-experts themselves trust the systems
- Explanations addressed explicitly to BB-experts as a support to their understanding and "debugging" of the systems should be considered as a serious research topic
- In particular explaining (and then preventing) gross errors might be more important than explaining correct behaviors
- And the fact that not everything can be explainable should be accepted as a reality

Some layman suggestions: assessing BB vs expert explanations

- Also human domain experts may not always provide convincing explanations
- It would be interesting (maybe it has already been done?) to give the same (previously unseen) cases both to some domain experts and to some BB systems and to compare not only the outcomes but also the provided explanations
- Would the explanations of domain experts be coherent? (assuming the outcomes are coherent)
- How would they compare with machine-produced explanations?

Some layman suggestions: stupidity-freeness vs. "intelligence"

- Explaining the construction, training, testing and debugging of a system may be achievable and support user trust even if the core behavior of the system remains black-box
- Evaluation of black-box systems might/should include some specific activity devoted to detection of potential gross errors
- Stupidity-free rather than (just) intelligent systems might be more trustable even if "unexplainable"

I'M NOT STUPID.
AND SOMETIMES I THINK
THAT'S PART OF THE PROBLEM.