WHAT DO MACHINES THINK ABOUT?

Gabriel Marín Díaz, Ramón A. Carrasco González, Daniel Gómez González
Universidad Complutense de Madrid (Spain)
gabriel.marin@ucm.es; ramoncar@ucm.es; dagomez@estad.ucm.es

EXTENDED ABSTRACT

Can machines think? This question was posed by Alan M. Turing (1950) in the mid-20th century. The answer to that question is the proposal of the so-called Turing test. In this test, **Artificial Intelligence** (AI) is considered to be a way of acting that imitates the intelligent behavior of human beings. Since then the AI has been surpassing the human being in tasks for which it was supposed to have intelligence: strategy games like chess, driving vehicles, composing symphonies, automatic planning, and a long etcetera that seems to have no end. In fact, the changes produced in the last decades in the telecommunications sector, accompanied by the development of the storage and processing capacity of information have meant a change of paradigm to which the name **Industry 4.0** has been given.

AI corresponds to a field of knowledge that includes **Machine Learning** (ML) and **Deep Learning** (DL). In both fields, to solve a problem proceeds to the training of models to learn the problem in question from existing data. Once the rules are obtained, we can apply them to new data sets to produce the appropriate answers by applying the rules learned from experience. To perform ML processes at least three fundamental parts are necessary: input data, the expected results and the measurement of the algorithm's performance so that the algorithm's work can be adjusted by means of feedback processes (Casella et al., 2013).

Interpretability in Machine & Deep Learning processes

An ML model, once implemented, can complete a task much faster and more reliably than any human, delivers consistent results reliably, and can be infinitely replicated. Training a person to perform a task with the same efficiency is costly and can take years.

An important aspect of using the ML is the interpretability of the models once they have been trained. From this point of view some authors distinguish two types of models (Liu et al., 2016):

- **White box models**, are models whose predictive or pattern identification behavior can be clearly explained based on the variables involved. Therefore, it is relatively simple to investigate the rules that such models have inferred from the data.
- **Black box models**, are models whose rules are not understandable in a simple way for the human being, it would be very difficult to explain how the system came to take a concrete decision before a certain entry (Liu et al., 2016).

Some authors even question the interpretability of white box algorithms (Lipton, 2018). In Figure 1 it can be seen, as a general rule, that the greater the interpretability of the ML algorithm, the lower its degree of flexibility and therefore the lower its degree of reliability. In other words, at present there is no doubt that the most powerful algorithms are not interpretable.

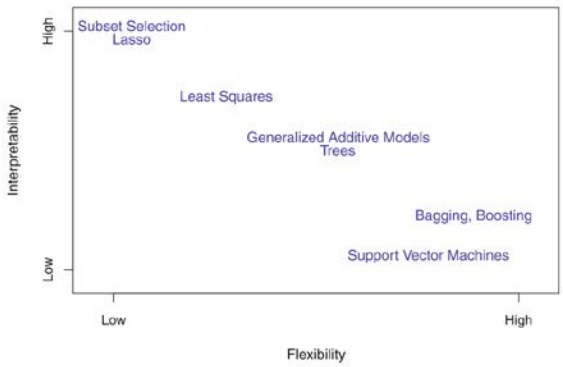

Figure 1. Interpretability vs. Flexibility of ML algorithms (Casella et al., 2013)

Therefore, seven decades after Turing posed his philosophical question, we could ask ourselves today the following question: What do machines think about? The question is that in some cases it may not be relevant to understand why a certain decision has been made, especially in low-risk environments. Although, in most cases the human being should understand why a decision that affects him individually or as a group has been taken. Examples include: the decision to grant a loan, a medical decision, driving a car, a selection process for a certain job...

In a study conducted in 2019 by Brandon Fornwalt, Geisinger Medical Center, Pennsylvania, they trained two AI algorithms capable of predicting the risk of death in the first year through the reading of electrocardiograms, even of apparently "normal" people, the accuracy of the algorithm was 85%.

The bias and its impact

ML models have been shown to learn very well from the data, but they also collect biases that may or may not be incorporated into the training data. This could make the training model potentially sectarian and discriminate against certain individuals. These potential biases constitute a fundamental point in the investigation of the problem of their interpretability (Miller, 2019; Molnar, 2019). There are three types of classical biases: statistical bias is determined by data collection, its origin, cultural bias derives from the language we speak, how we express ourselves, our stereotypes, and cognitive bias identifies our beliefs and values.

Below we will list some examples of bias applied by ML algorithms in different technologies and areas.

- In 2019, Ruby on Rails entrepreneur David Heinemeier Hansson shared a disturbing story on Twitter, claiming that the Apple card was discriminating against his wife.
- In 2015, the Google Photos application labeled two African Americans as "gorillas". Google engineers analyzed the account and found that the algorithm had trouble adjusting to the photo contrast, lighting and skin tone.
- In 2016, it was noted that some of the LinkedIn algorithms were gender biased, recommending better paying jobs to men.
- In 2016 Microsoft launched "Tay," a chatbot that was intended to mimic the behavior of a curious teenage girl seeking to engage in informal conversation at RRSS. Within 24 hours, Tay was tweeting her empathy for Hitler or her support for the genocide.

Bias in the source data implies biased decision making, and if the algorithm used is black box, it will be much more difficult to identify such bias.

Conclusions

So, to answer the question, what do machines think about? The answer has to be approached from different points of view, several areas of action are listed to answer this question:

1. Scientific Vision

Today there are various scientific approaches that aim to help explain in a relatively simple way the models of ML, especially the so-called black box models. These interpretation methods can be classified according to several criteria (Lipton, 2018; Molnar, 2019; Ribeiro, 2016): intrinsic, post hoc, specific, agnostic, local and global.

These approaches, although interesting, are far from turning what black box algorithms do into a white box. Therefore, new actions in the scientific field can be expected for the problem posed.

2. Legislative Vision

Profiling and automated decisions can pose significant risks to individual rights and freedoms. The European and Spanish legislation on data protection obliges and requires certain guarantees. Article 22° of the General Regulation on Data Protection (RGPD) establishes that European citizens have the right not to be the object of a decision based solely on automated means, including the elaboration of profiles. As it usually happens in areas of Information and Communication Technologies (ICT), the legislative aspects go behind the technological advances, so it could be questioned if this regulation can be enforced in a global and effective way. Furthermore, if it is enforced, can the RGPD regulations really protect us from decisions made by an algorithmic bias?

3. Performance of Independent Entities

It seems clear that there is a need for action by independent bodies capable of determining the "quality" of the algorithm, providing sufficient guarantees to citizens, thus increasing social acceptance of this type of practice. Ensuring that the following qualities are met: fairness, privacy, reliability, soundness, causality, trust (Doshi-Velez, 2017).

As a result of this need, in December 2019 the technical subcommittee for standardization CTN 71/sc 42 – Artificial Intelligence and Big Data was set up in Spain precisely to draw up standards in the field of AI, participating in the development of the standards at a global level that are being developed in the international committee ISO/IEC JTC 1/SC 42 Artificial Intelligence.

These actions, of course, are very limited and timid at present.

4. Business Vision

Companies, especially those in the ICT sector, have undertaken certain actions on their own to address the problem.

- IBM launched in 2018 the Fairness 360 Kit project (IBM, 2018), this toolkit helps to examine, report and mitigate discrimination and bias in ML models.
- Microsoft has a model interpretation SDK in Azure ML Python package (Microsoft, 2020).
- Google has an Explainable API (Google, 2020), which is a set of tools and frameworks capable of helping to debug and understand the behavior of ML models.

Other companies are taking steps in the same direction, although it is obvious that companies will put their own interests before the problem posed.

In short, when faced with the question of what machines think, steps are being taken, as we mentioned in the previous points. However, the question that should concern us most is whether this question is really asked by today's society. In a society that is becoming more and more Innumeracy (John A. Paulos, 1988), where the decision making is usually made through System 1 of thought: fast, intuitive and emotional (Daniel Kahneman, 2011), as opposed to System 2 of thought: slow, deliberative and logical. We can intuit that this speed, immediacy of the daily and short term could come to invade us, leaving the decisions that require thinking and meditating to a third party, the machine, which also would not have to explain the reason for its decisions.

KEYWORDS: Machine Learning, Interpretability, Deep Learning, Bias, Artificial Intelligence

REFERENCES

Casella, G., Fienberg, S., & Olkin, I. (2013). An Introduction to Statistical Learning. In Springer Texts in Statistics. http://books.google.com/books?id=9tv0taI8l6YC

Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. Ml, 1–13. http://arxiv.org/abs/1702.08608

Lipton, Z. C. (2018). The mythos of model interpretability. Communications of the ACM, 61(10), 35–43. https://doi.org/10.1145/3233231

Liu, H., Cocea, M., & Gegov, A. (2016). Interpretability of computational models for sentiment analysis. Studies in Computational Intelligence, 639(March), 199–220. https://doi.org/10.1007/978-3-319-30319-2_9

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Molnar, C. (2019). Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. Book, 247. https://christophm.github.io/interpretable-ml-book

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Model-Agnostic Interpretability of Machine Learning. Whi. http://arxiv.org/abs/1606.05386

John A. Paulos – Innumeracy: Mathematical Illiteracy and its Consequences. Hill and Wang. 1988. ISBN 978-0-670-83008-4

Kahneman, D. (2011) Thinking, Fast and Slow, Farrar, Straus and Giroux, ISBN 978-0374275631

Turing, A. M. (1950). Computing Machinery and Intelligence. Mind 49, 433-460

IBM Fairness 360 Kit (2018), https://aif360.mybluemix.net/

Microsoft Azure ML Python (2020), https://docs.microsoft.com/es-es/python/api/overview/azure/ml/

Google Explainable API (2020), https://cloud.google.com/explainable-ai