Metadata

- Id: EU.AI4T.O1.M4.2.1t
- Title: 4.2.1 Ready to trust AI for decision making?
- Type: text
- Description: Understand the impact of using decision-making AI tools and the necessary precaution of use
- Subject: Artificial Intelligence for and by Teachers
- Authors:
  - AI4T
  - AI4T
  - Ikram Chraibi Kaadoud, AI Researcher
- Licence: CC BY 4.0
- Date: 2022-11-15

ARE YOU READY TO TRUST AI FOR DECISION MAKING?

All decisions made with AI-based tools do not have all the same impact.

For some automated decisions, such as the "solution steps" suggested to a student by a mathematical problem-solving application, the long-term risk and harm can be *considered* rather low.

Other decisions, on the contrary, present a potential harm and/or risk.

There is then a maximum of precautions to be taken and first of all the decision must be explainable (why this decision is proposed for this particular situation, for this particular student or group of students).

Let's see some criteria that are used to "evaluate" the decision-making process of AI-based systems.

EXPLAINABILITY

Explainability - one of the 7 key requirements for trustworthy AI: "*Explainability concerns the ability to explain both the technical processes of an AI system and the related human decisions (e.g. application areas of a system). Technical explainability requires that the decisions made by an AI system can be understood and traced by human beings.*"[1]

In the field of education, this means that in any AI decision-making tool, the way a decision is proposed and the human interaction involved are elements that need to be accessed.

This requirement is more or less easy to meet, but for some AI technologies the explainability is not so simple to provide. For example, in the case of neural networks with many layers, explanations can be difficult to render. So that a new area of AI is now developing: eXplainable AI or XAI defined as an *artificial intelligence in which humans can understand the decisions or predictions made by the AI. It contrasts with the "black box" concept in machine learning where even its designers cannot explain why an AI arrived at a specific decision."*[2]

INTERPRETABILITY

Predictions made with some AI techniques are easier to interpret than others. A prediction made from a decision tree, for example, is explainable. But it is not always the most interesting predictions to be made.

At the opposite end of the explainability spectrum, there is Deep learning, which can be difficult to explain but whose output may be much more significant than those made with highly explainable AI.

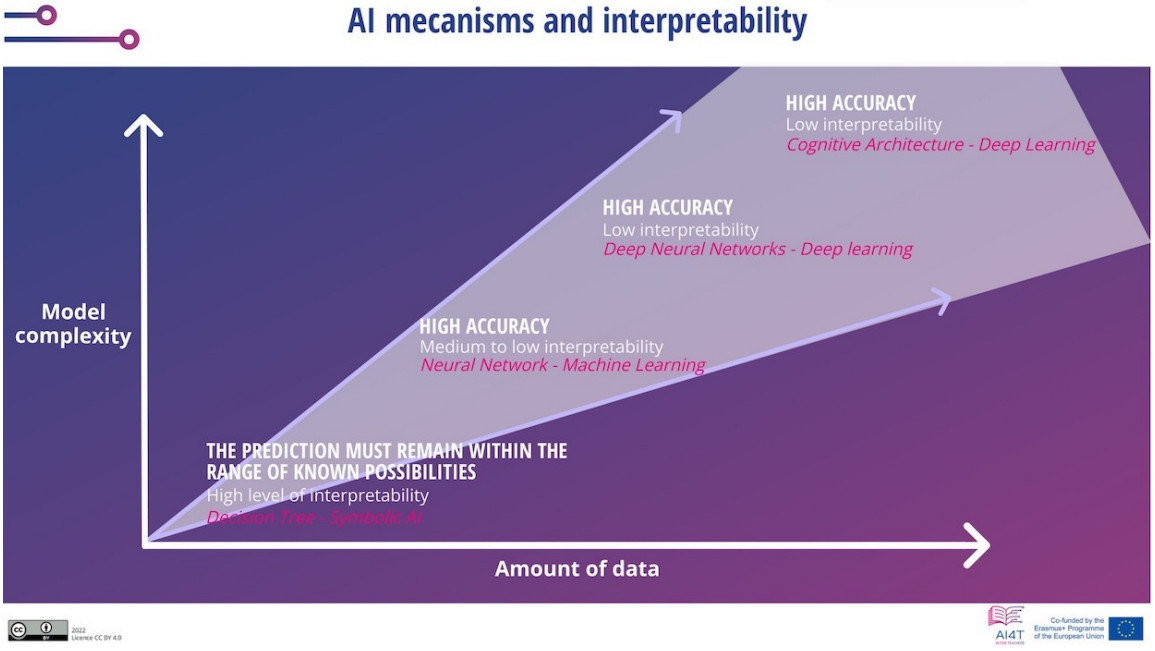

Figure1: AI mechanisms and interpretability. Adapted from Mooc IAI / Ikram Chraibi Kaadoud - CC.BY.SA 2.0.

Thus, the decision support provided by tools with low interpretability can be more significant than the support provided by tools with high interpretability.

FROM DESCRIPTION TO PRESCRIPTION

Here is a representation that relates the technology used, its complexity and its strategic outcome.

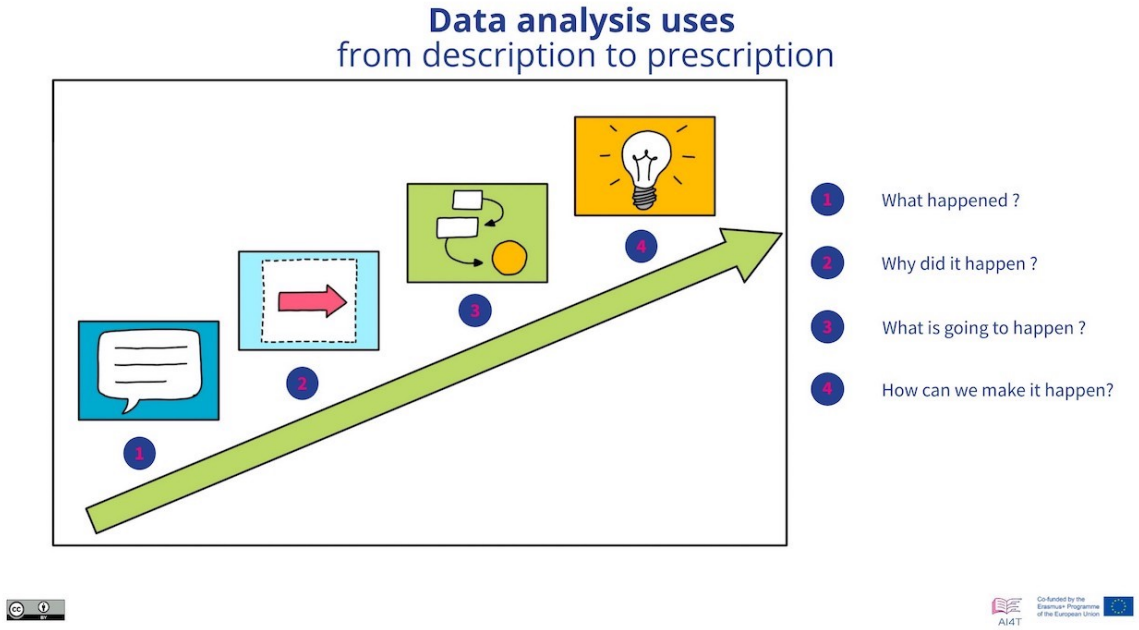

Figure 2: Classification of data analysis use from description to prescription[3] (Adapted from the "Learning Analytics" video of this course).

In the following 4 categories the correlation between the complexity of the methods used and the strategic results can be noted.

DESCRIPTIVE ANALYTICS

Descriptive Analytics looks into data to answer the question "What happened?". Those can be provided in the form of "*simple summaries about the sample and about the observations that have been made. Such summaries may be either quantitative or visual, i.e. simple-to-understand graphs*"[4]. It is based on traditional tools without AI.

DIAGNOSTIC ANALYTICS

Diagnostic Analytics answers the question, "Why did it happen?". It leads to the identification of the nature and cause of a phenomenon to determine the mitigations, and solutions. Some techniques used for diagnostic analytics: Statistics methods like data discovery, data mining and correlations. Those methods can imply AI.

PREDICTIVE ANALYTICS

Predictive Analytics examines data or events to answer the question "What is going to happen?" or more precisely, "What is likely to happen?". "*Predictive analytics is forward-*

looking, using past events to anticipate the future. Predictive analytics statistical techniques include data modelling, machine learning, AI, deep learning algorithms and data mining."[5]

PRESCRIPTIVE ANALYTICS

Prescriptive Analytics answers the question "What should be done?" or "How can we make it happen?".

"*Prescriptive analytics not only anticipates what will happen and when it will happen, but also why it will happen. Further, prescriptive analytics suggests decision options on how to take advantage of a future opportunity or mitigate a future risk and shows the implication of each decision option*."[6]

In summary, the more relevant the tools can be as an aid to decision making, the more complex the information technologies are and the more difficult they can be to explain. But in terms of the assistance provided, attention must be maintained on explicability and any vigilance required in the use of the IA tool in an area where the consequences of decisions are significant and lasting.

---

1. "*Moreover, trade-offs might have to be made between enhancing a system's explainability (which may reduce its accuracy) or increasing its accuracy (at the cost of explainability). Whenever an AI system has a significant impact on people's lives, it should be possible to demand a suitable explanation of the AI system's decision-making process. Such explanation should be timely and adapted to the expertise of the stakeholder concerned (e.g. layperson, regulator or researcher). In addition, explanations of the degree to which an AI system influences and shapes the organisational decision-making process, design choices of the system, and the rationale for deploying it, should be available (hence ensuring business model transparency).*" From "Ethics Guidelines for Trustworthy AI" (consulted 10/16/2022). ↩
2. From wikipedia article on "Explainable artificial intelligence" (consulted 10/16/2022). ↩
3. See in this course the section 1.1.3. on Learning analytics (video). ↩
4. From wikipedia article on "Descriptive statistics" (consulted 10/16/2022). ↩
5. From wikipedia article on "Predictive Analytics" (consulted 10/16/2022). ↩
6. From wikipedia article on "Prescriptive Analytics" (consulted 10/16/2022). ↩