EXPLAINABLE MACHINE LEARNING FOR ETHICAL ARTIFICIAL INTELLIGENCE BASED DECISIONS

Radu STEFAN[1]
George CARUTASU[2]

Abstract: *In the last century, many approaches and tools have been developed to implement systems that would achieve Artificial Intelligence (AI) and solve most difficult problems in computer science with that. The quest is to find the most suitable machine learning method for a given problem domain. Currently the approach that yields the best results in practice in most domains is based upon artificial neural networks, or the more advanced deep neural networks. Neural networks are highly efficient for most common scenarios, e.g. language understanding, image recognition and the like. Typically, models are trained for a specific task, and their performance is judged based on the binary outcome, e.g. in image recognition, whether the artefact was recognized or not. From an ethical decision-making point of view, the challenge remains to identify how the learning process influences the outcome and finally to rationally understand how a decision has been made. In this paper we present a framework for an enhanced machine learning model, that adds a layer of decision explanation to the outcome towards the user of the system. We suggest how the learning model needs to be expanded to an explainable model and the corresponding explainable interface that would allow presentation towards the user.*

1. Introduction

There is a problem with artificial intelligence. It can be amazing at churning through gigantic amounts of data to solve challenges that humans struggle with. But understanding how it makes its decisions is often very difficult to do, if not impossible. That means when an AI model works it is not as easy as it should be to understand why they're doing what they're doing. In order to achieve ethical decision making, the prerequisite consist in the ethical dimensions framework we have detailed in an earlier publication called *"How to approach ethics in intelligent decision support systems"*. For at least two of the ethical dimensions, that we would want to address here, it is crucial to have the ability to understand whether the system operates correctly and to understand how decisions are taken:

- Reliability & Safety, and
- Transparency

[1] Dipl.-Ing. Radu Stefan (Politehnica University of Timisoara, UPT), radu.stefan@live.de

[2] Prof. PhD George Carutasu (Politehnica University of Timisoara, UPT – Romania-American University, URA), carutasu.george@profesor.rau.ro

2. The need for Reliability, Safety and Transparency of artificial intelligence based decisions.

Reliability, Safety and Transparency of artificial intelligence based system, typical address the understanding of how a system operates and whether it operates correctly up to the understanding of how decisions are taken.

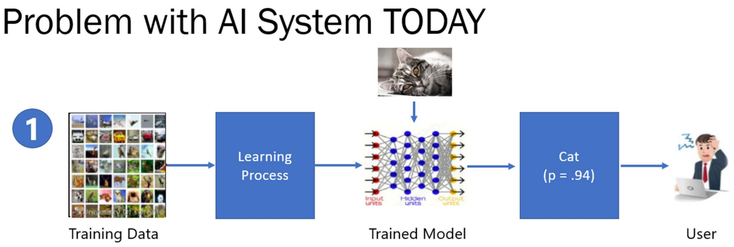

Problem with AI System TODAY

Figure 6 - User output: Cat (p = .94)

2.1. Reliability

By reliability we refer to a system that is consistent and accurate. On contrast unreliable systems are typically inconsistent and not accurate, thus mostly non-usable. For AI based systems the challenge in reliability, so in consistency and accuracy, lies in the very nature of probabilistic approaches. While for some domains an accuracy of 99% of producing the correct outcome may be considered accurate (e.g. placing ads based on user profile), some other domains (such as automotive) may require a 99.9997% accuracy.

Artificial intelligence systems, must operate at the domain specific accuracy, where accuracy is defined as the proximity of the outcome to the correct value. We intentionally assume accuracy as the key factor for reliability vs. precision that would be more relevant for the internal operation and not the outcome of the system for the end-user. The actual difference between *reliability* and *precision* is depicted in the image below.

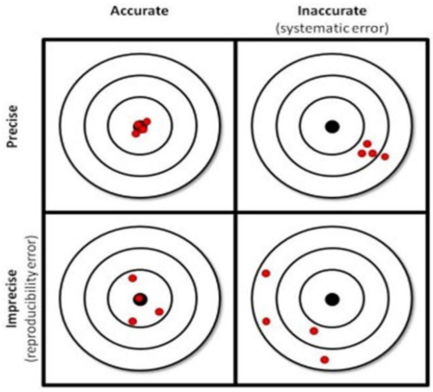


Figure 7 - Accuracy vs Precision

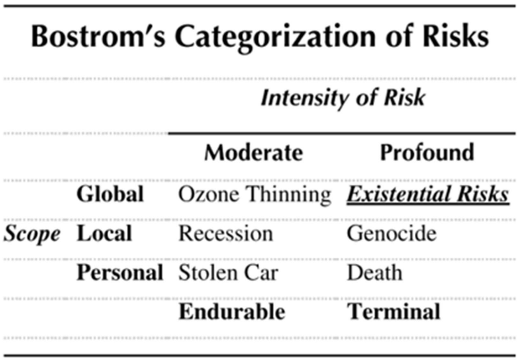
2.2. Safety

Safety, in contrast to reliability, refers to a system that operates in a manner which does not hard others, directly or indirectly. This is valid for a system action as well as for an in-action that would conduct to harm (especially harming humans).

Like the reliability, the safety of an AI based system needs to be defined in the context and the environment such a system operates. For example, AI based systems that operate a self-driving car, or such that are used in patient diagnosis are, amongst others, highly sensitive to safety of humans. This is also valid for most of embodied AI robots, as those systems have a physical degree of freedom, thus hypothetically capable of harming humans. On the other end of the spectrum would be AI systems that operate solely as a software system, e.g. recommender system for online shops or e-mail spam filtering. In general, those systems do not pose any risk to the safety of humans.

However, looking at a common example in artificial intelligence based systems, such as a credit scoring application used by financial institutions, the safety aspect would be harming humans that would be rejected a load, although there is would not be an objective reason to do so. There are examples where credit scoring application are gender or race biased, even if the gender was eliminated as an input. For example, gender bias can be derived from written text as shown by [1].

A more general approach to classifying risks was formulated by Bostrom [2].



Bostrom's Categorization of Risks

		Intensity of Risk	
		Moderate	**Profound**
	Global	Ozone Thinning	***Existential Risks***
Scope	**Local**	Recession	Genocide
	Personal	Stolen Car	Death
		Endurable	**Terminal**

2.3. Transparency

Transparency of artificial intelligence systems, as the third aspect in this chapter, refers to the ability of the user to understand the outcome of the system. In this sense we look at the interpretability of the results. This relevant in the same manner for both levels of users: *novice* and *expert.* Typically AI based system are used in one of the two scenarios: *work* or *learn.* In the learn scenario we consider any user a novice, on contrast to the working (or usage) scenario.

(Domain) Experts using AI bases system, typically do have some types of knowledge, like: *terminological* and *strategic knowledge,* however they might miss the *justification* or *trace knowledge.* For novice users it is expected that they do lack any of the four types of knowledge, represented in the table below [3] :

Type of Knowledge	Description and Purpose	Illustration of question requesting explanations using such knowledge
Terminological knowledge Synonyms: definition knowledge	Knowledge of concepts and relationships of a domain that domain experts use to communicate with each other. In order for one to understand a domain, one must understand the terms used to describe the domain.	What is the definition of gross domestic product?
Justification knowledge Synonyms: Why, descriptive knowledge	"Textbook rudiments" which are required before one can solve problems. Justification knowledge provides abstract factual knowledge about a domain, typically represented declaratively.	Why is inflation dependent on the money supply?
Trace knowledge Synonyms: How, problem solving knowledge.	Knowledge about how tasks have, or are about to be accomplished.	How did you conclude that the patient has diabetes?
Strategic knowledge Synonyms: Control knowledge	Knowledge about the system's control behaviour and problem solving strategy.	Why do you need to know if the patient has ever had mumps?

Table 2 - Types of Knowledge

3. Scope of Explainable Artificial Intelligence (XAI)

In order to address the shortcoming in Reliability, Safety and mostly in the transparency of AI based systems, an additional layer of *explainability* is required for every system. This layer shall be an integral part of the model and needs to be constructed along the way with the AI model itself.

The scope of the *Explainable Artificial Intelligence* lies at the intersection of three major areas:

- Artificial Intelligence
- Social Science, and
- Human-Computer Interaction

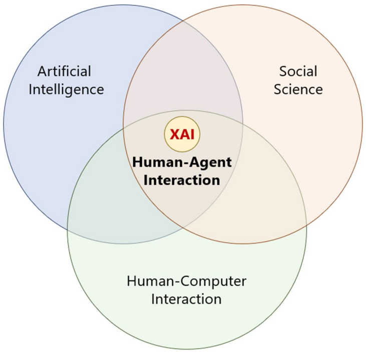


Figure 8 - Scope of Explainable Artificial Intelligence

Behind the *Artificial Intelligence* domain there is a specific machine learning model, that powers the system and produces on outcome based on the input. For systems developed in the last decades, most machine learning models produce results based on probabilistic models, therefore the outcome also has a probability attached. It is fairly known, that humans not trained in statistics are inclined to interpret probability wrongly, which goes both ways. A reasonable probability maybe mistrusted, or a poor probability might be accepted as a good outcome, depending on the understanding and the believe as defined in the social science.

The *Social Science* domain is concerned with human perception. Heider [4] defines social attribution as person perception. A detailed analysis would go beyond the scope of this paper, for the XAI model however it needs to be noted that human are in need of an explanation or a reason to believe, which is necessity to trust.

Lastly, the domain of *Human-Computer Interaction (HCI)* is crucial for generation a good explanation, as HCI defines how information is presented to the user and how the users can act upon the computer system. The form of interaction can significantly influence the perception of the user.

Overall an *Explainable Artificial Intelligence* system, might have several beneficial outcomes, that could address the needs presented in *chapter 2 The need for Reliability, Safety and Transparency of artificial intelligence based decisions.*

3.1. XAI to increase reliability and enhance learning

In scenarios where artificial intelligence systems are used for learning and training, the explanatory part is necessary. The ability to learn is strongly connect with the ability to reason, thus systems used in learning scenarios should present explanations in at least all four types of knowledge as described in *Table 2 - Types of Knowledge.*

3.1.1. Terminological Knowledge

Every AI system should be accompanied by a terminological knowledge base that defines concepts and relationships of the particular domain. Based on keywords or other means of interrogation, users (or learners) should be able to interrogate specific terminology.

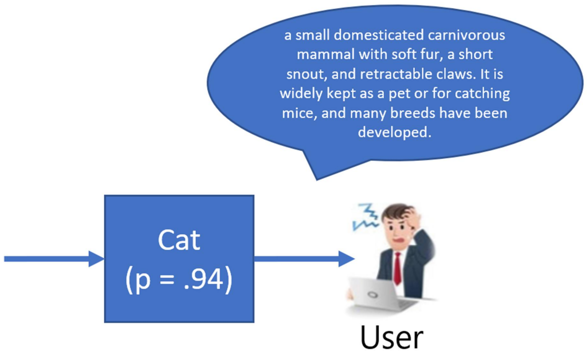
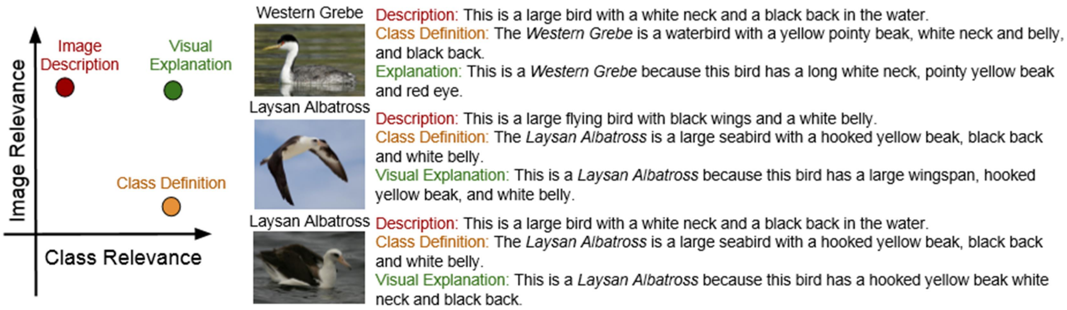

a small domesticated carnivorous mammal with soft fur, a short snout, and retractable claws. It is widely kept as a pet or for catching mice, and many breeds have been developed.
Cat
(p = .94)
User

Figure 9 - Additional Terminological Knowledge

A superior functionality would be given, when the system automatically triggers explanations based on the given outcome and presents them to the user. The *Figure 9 - Additional Terminological Knowledge* shows a schematic example.

3.1.2. Justification Knowledge

A secondary addition to explanation is given by the *justification knowledge (or descriptive knowledge).* A good example for justification knowledge especially in visual interpretations is given by Hendricks [5]:



Description: This is a large bird with a white neck and a black back in the water.
Class Definition: The *Western Grebe* is a waterbird with a yellow pointy beak, white neck and belly, and black back.
Explanation: This is a *Western Grebe* because this bird has a long white neck, pointy yellow beak and red eye.

Description: This is a large flying bird with black wings and a white belly.
Class Definition: The *Laysan Albatross* is a large seabird with a hooked yellow beak, black back and white belly.
Visual Explanation: This is a *Laysan Albatross* because this bird has a large wingspan, hooked yellow beak, and white belly.

Description: This is a large bird with a white neck and a black back in the water.
Class Definition: The *Laysan Albatross* is a large seabird with a hooked yellow beak, black back and white belly.
Visual Explanation: This is a *Laysan Albatross* because this bird has a hooked yellow beak white neck and black back.

Figure 10 - Class vs Image Relevance

Using a *Description* and a *Class Definition* (from the terminological knowledge) an explanation can be generated.

3.1.3. Trace Knowledge

In image recognition a simple model to realize trace knowledge is by recognizing sub elements of a picture. This can be approached top-down or bottom-up. The first would be to recognize the image as a whole and classify it. In a second iteration to lookup in the knowledge base for sub elements and lastly to rerun image recognizer to find the sub elements in the image. Using the cat example, it would mean:

I. Image recognizer -> cat (p = .94)
II. Look up cat class for sub elements -> fur, claws, etc.
III. Rerun imaged recognizer to identify fur, claws, etc. -> if true -> traceable!

Another approach would be to reverse the above order and to run the algorithm bottom-up. First recognize sub elements of the image, second search through knowledge what suits most all the sub elements, and lastly rerun image recognizer to ensure result from second step is confirmed.

Which of the two models is most appropriate, depends largely on the existing knowledge and the problem domain and on the model selected. Choosing the best model based on those dependency should be further researched in a separate paper.

3.1.4. Strategic Knowledge

4. The fourth knowledge type is mostly about the system and the operations that lead to the result present to the user. In this step a description (understandable by a novice and domain uneducated human) should be presented. A cumulation of the three methods described previously would suffice to increase the learning effect.

5. By designing implementing the above knowledge types into an AI system, not only the learning / usage experience would be enhanced, but implicitly the reliability of the system would be significantly increased. Especially the *justification* and the *trance knowledge* force the system into a higher reliability.

5.1. XAI to increase safety by avoiding (more) mistakes

An artificial intelligence system, that would incorporate all four of the knowledge types mentioned in the previous chapter, would be less prone to mistakes and errors. A terminology knowledge system would force designers and implementers to develop the system more accurately, but the real benefit is with the human user, that has the possibility to judge the result in the context of the terminology and feedback his or her perception back to the system or the model.

The justification knowledge ads the same benefits as the terminology knowledge, however it is surely more complex to embedded and from a financial point of costlier for the entire system.

The trace knowledge is the most significant contributor to reducing mistakes and it implements a secondary step of ensuring the answer is correct. It is obvious that the accuracy is highly improved, however it depends on the domain of usage to what extend the trace knowledge is rather trivial to implement (as in the image recognition example) or highly complex.

The strategic knowledge also contributes to less errors, as it forces designers to precisely document and communicate the operations of the system.

5.2. XAI to increase transparency by enhancing trust

Lastly increasing transparency and thus enhancing trust is the major scope of *explainable artificial intelligence.* Trust is something that needs to be looked at from an end-user perspective, no matter if a novice or an expert user.

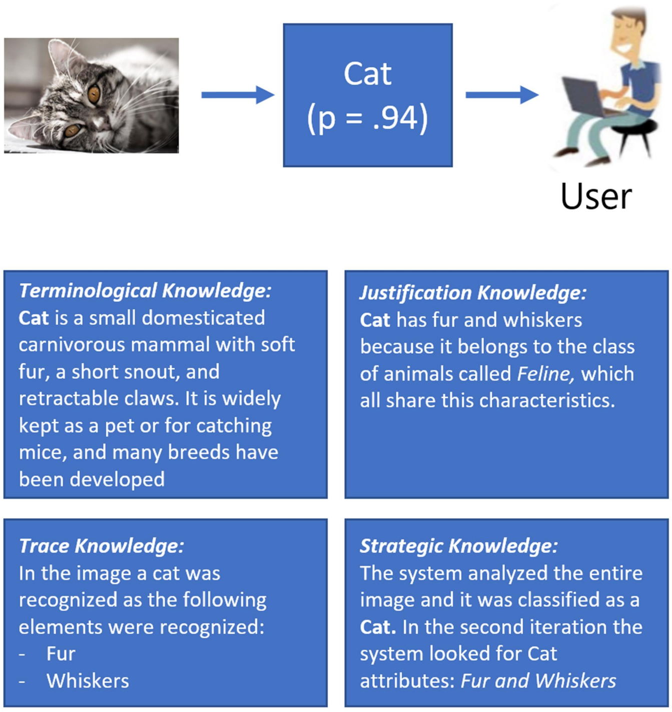
On top of the technical implementations discussed above, the transparency and the trust, related mostly to the *human-computer interaction (HCI)* domain, as they are highly dependable on the user's perceptions.

Interfaces need to be designed appropriately for the domain intended. For example, an interface for medical staff using an artificial intelligence system for diagnosis will differ significantly from an interface a consumer is using to sort family pictures from the last year's vacation across Europe.

6. Extension of current approach

By designing implementing the above knowledge types into an AI system, not only the learning / usage experience would be enhanced, but implicitly the reliability of the system would be significantly increased. Especially the *justification* and the *trance knowledge* force the system into a higher reliability.

As a general approach, all techniques described in this chapter could be incorporated in the user's screen. An example is depicted below for the same animal image recognizer.



Terminological Knowledge:
Cat is a small domesticated carnivorous mammal with soft fur, a short snout, and retractable claws. It is widely kept as a pet or for catching mice, and many breeds have been developed

Justification Knowledge:
Cat has fur and whiskers because it belongs to the class of animals called *Feline,* which all share this characteristics.

Trace Knowledge:
In the image a cat was recognized as the following elements were recognized:
- Fur
- Whiskers

Strategic Knowledge:
The system analyzed the entire image and it was classified as a **Cat.** In the second iteration the system looked for Cat attributes: *Fur and Whiskers*

Figure 11 - Additional Knowledge to be presented

7. Explainable AI in practice

For explainable artificial intelligence to work in practice, it is necessary to construct the knowledge system along with the artificial intelligence model itself. We call the knowledge system the *explanation model.*

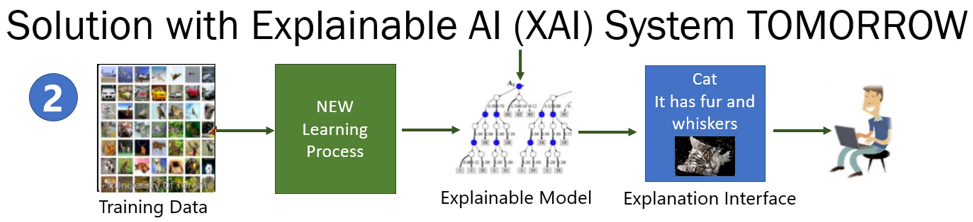
In a more general approach, all the above knowledges can be summarized as a justification narrative. In order to generate the narratives features of the data set need to be classified on importance and effect. Such a model is presented by Biran [6]:

Effect Importance	High positive	Low	High negative
High positive	Normal evidence	Missing evidence	Contrarian counter-evidence
Low	Exceptional evidence	Negligible	Exceptional counter-evidence
High negative	Contrarian evidence	Missing counter-evidence	Normal counter-evidence

Table 3 Narratives Importance vs Effect

Extracting the evidence, as positive or negative effect during the learning/training phase of the model and having it available during operations, gives a framework that allows for every new observation, to construct narratives using positive and negative statements.

Converting features into human readable text is a separate research and work area that would go beyond the scope of this paper.



Solution with Explainable AI (XAI) System TOMORROW

As a result, the learning process is extended to extract evidence and the outcome is embedded in the explainable model. This allows the user interface to be enriched with an explanation interface.

8. Conclusion

In conclusion we suggest an enhanced framework for explainable artificial intelligence based systems that includes ethical aspects of reliability, safe and most important transparency. Those aspects cover two major ethical dimensions as suggest by us in an earlier publication [7].

In future research, the approaches described in this paper need to be applied to specific domains and tested in practice. In some areas the applicability might be limited while in others many extensions might be possible.

The above implementations would contribute towards trustworthy AI, an initiative driven by the European Union. A final paper is expected to be published in May. 2020.

7. References

[1] A. Caliskan, "Semantics derived automatically from language corpora contain human-like biases," *Science, Vol. 356, Issue 6334,* pp. 183-186, 2017.

[2] R. V. Yampolskiy, Artificial Intelligence Safety and Security, Boca Raton, FL: CRC Press/Taylor & Francis Group, 2018.

[3] K. Darlington, "Aspects of Intelligent Systems Explannation", *Universal Journal of Control and Automation 1 (2),* pp. 40-51, 2013.

[4] F. Heider, The psychology of interpersonal relations, New York: Wiley, 1958.

[5] L. A. Hendricks, "Generating Visual Explanations," in *Computer Vision -- ECCV 2016*, Cham, Springer International Publishing, 2016, pp. 3-19.

[6] O. Biran and K. McKeown, "Justification Narratives for Individual Classifications," *ICML 2014 AutoML Workshop,* 2014.

[7] R. Stefan and G. Carutasu, "How to apporach ethics in intelligent decision support systems," *Innovation in sustainable Management and Entrepreneurship,* 2019.