CM4107 Coursework #2

Anthony Sébert

School of Computing Science and Digital Media RGU Aberdeen, UK

Abstract

This paper aims to address the problem of explainable systems in artificial intelligence. Since the domain's applications become more numerous and present in people's lives as the technology matures, the need for explainable systems is emerging, for both ethical and legal concerns. Of course, it is not possible nor desirable to deploy on a large scale a product based on a technology that is not fully understood, especially in areas where a high level of reliability and confidence is necessary. For that purpose, we first conduct a comparative study on three machine learning algorithms using the Python language and the ScikitLearn library. Then we define and address the aims of the concept of Explainable Artificial Intelligence, and finally, we discuss the ethical concerns about Artificial Intelligence in general, in connection with existing ethics philosophies.

ACM classifiers

• **Computing methodologies~Artificial intelligence** • Computing methodologies~Machine learning algorithms • Social and professional topics~Codes of ethics

Comparative Study using ScikitLearn

Unfortunately this section could not be covered due to lack of time.

CART

k-Nearest Neighbors search with Ball Tree

Multilayer Perceptron with Stochastic Gradient Descent

Explainable AI (XAI)

Overview of explainable XAI

Explainable AI are artificial systems that mimic human intelligence and are based on transparency, a highly desirable characteristic by the way, so as to make the decisions and the processes that lead to those understandable by humans [1]. It is on the opposite of the black box approach of artificial systems such as traditional machine learning algorithms, that moreover does not ensure trust towards AI.

In order to reach a satisfactory level of transparency to explain and predict the operation of the AI, either it is right or wrong (because explain an error will always more valuable than explaining a success).

Over the past years and the growth of the AI presence in the society, regulators, official bodies and general users have developed a legitimate need for understanding AI. So if the artificial system is meant to make decisions that impact the civil society, it is a legal requirement to provide an explanation of the decisions, especially in a context where AI is known to reproduce human bias [2] and the public concern for new technologies and data protection arise (EU GDPR).

The *Five Goals of Explainability* sum up all the challenges of the domain [3]:

- Transparency
- Justification
- Conceptualization
- Relevance
- Learning

Legal, ethical and Social implications of XAI

Since the artificial neural networks, the most common solutions to solve complex artificial intelligence problems, currently operate as black boxes, the main challenge is to invent or adapt algorithms and frameworks that enforce transparency by making the decision process more understandable and predictable. Even if progress has been made to "demystify" their operation [4] [5], the existence of a gap in theoretical research concerning the behavior of ANN is a serious drawback to their ethical utilization.

This is an even larger concern when XAI is employed in sensitive fields such as:

- Autonomous vehicles
- Algorithmic trading
- Medical diagnoses

In every case, the system must also be able to explain a decision afterward, in order to provide a key of comprehension in the event of a problem or concern about the decision [6]. Further research is also needed in that particular area to create common explanable model or model architecture.

Explainability and Transparency on chosen algorithms

CART

The CART algorithm produces either classification or regression trees, depending on whether the dependent variable. The trees are formed by a set of rules based on the dataset model [7]. It is mainly used in data mining.

Advantages

- Based on simple conditional statements, they are easily understandable, even by non-technicians. A graphical representation might also be used to help the comprehension
- Close to the human decision-making process, which makes CART more transparent to inspection and potentially able to model human interactions
- A decision from a situation can be fully justified by the internal boolean logic behind it
- Can be demonstrated with statistical tests, a way to ensure the reliability of the system

Limitations

- Decision trees are very sensitive to changes, especially near the root

- Learners can create over-complex trees, really hard to read and understand for a human

Since its invention, CART has been modified or adapted to fit in specific purpose or remove/reduce its drawbacks [8] [9].

k-Nearest Neighbors search with Ball Tree

The Ball tree aims to solve the problems of the K-D tree algorithm, itself meant to find a better approach than brute-force [10]. It organizes points in a multi-dimensional space, partitioning them into hyperspheres ("balls").

In k-NN classification, the output is a class membership. An element is classified regarding the most common class among its k nearest neighbors [11].

Advantages

- The graphical representation is easy to understand and visualize, non-experts included

Limitations

- The validation of the results with a confusion matrix or statistical methods is not trivial to represent [12]
- Explainability might not be an evidence if the dataset is too sparse (too few neighbors)

Multilayer Perceptron with Stochastic Gradient Descent

MLP is a class of feedforward artificial neural network, constituted of layers of nodes (input layer, hidden layer(s), output layer) [13]. It is widely used in machine learning, especially in deep learning; backpropagation makes MLP a very good choice.

Advantages

- Transparent neural networks are yet to discover (open research field), progress has been made towards this goal [14] [15]

Limitations

- The learning process is opaque by nature, computation or modification of the model is necessary to get more understandable results
- It is impossible to know the classifier's confidence about the results since neural networks are not probabilistic
- Artificial neural networks structure is similar to biologic neural networks structure, but their operation is not the same, even if bridges can be found [16].

Ethical AI

Foreword

This section aims to address the topic of Ethical AI in its entirety. So as to ensure a good comprehension between the author and the reader, several concepts are to define.

Artificial Intelligence

A concise definition of AI could be the following:

> The ensemble of theories and techniques implemented so as to realize machines able to simulate intelligence.

It encompasses theoretical and practical works, the tangible means and the key goal, intelligence. But of course, many other definitions exist and are relevant as well. In fact, it seems that every professional of the sector sees its own "flavor" of AI.

Ethics

General definitions

The ethics (from the ancient Greek *ἦθος*: "accustomed place, custom, habit") is the branch of the philosophy studying the value judgments. It can be viewed as the basis of morality, its underneath and primal reflexion. Ethics are different from the scientific method which relies on fact judgments expressed in descriptive statements. For philosophers such as Aristotle and Kant, ethics is about defining *what needs to be*.

Consequentialism

The consequences of an action are an important aspect of it (it is the basis of experimentation, important in the learning process). These consequences can then be considered as relevant criteria to applies norms to behaviors. In that case, a decision is considered as good if the ensuing repercussions are a benefit. The ground of the evaluation is moved to the observable world rather than the system's internal logic. But that means that in order to state whether the actions of such systems are good or bad, we first need to define was is beneficial or not [17]. For example, if an AI drives a car, and by a traffic, hazard have to choose whether to kill a child or an elderly, what to decide?

Eudemonism

Eudemonism (from the ancient Greek *εὐδαιμονία*: beatitude) state that happiness (different from pleasure) the goal of (human) life. Therefore the ultimate criteria of choice in actions is happiness, towards the individual and/or the whole society. Eudemonism is based on a general trust in the human being. The doctrine focuses on this only chance of fulfillment that is earthly life and therefore it is the success of this life, immediate happiness or rationalized over a long time, both his own and that of others, that it consecrates logically the essence of its effort [18].

Deontological ethics

Kantian ethics has been described as deontological, that is to say, it considers action in itself and duty or moral obligation, independently of any empirical circumstance of action. It therefore also opposes the consequentialism, which estimates the moral value of the action according to the foreseeable consequences thereof. Because of the absolute imperative nature of the notion of duty, and the unnecessary connection between happiness and morality, the Kantian position has often been described as rigorous [19] [20].

Link with AI

When it comes to AI, a recent technological breakthrough (which was not even possible a few decades ago), made to be responsible for critical systems (anticipation, even if such real-life examples have already appeared), the potential impacts are not to minimize.

AI problems sorted by their causes

Whether in the fiction or in reality, numerous examples of AI harm have been provided over the past century. While AI misbehavior can sometimes be funny [21], some people have met their death sooner than expected [22]. Should intelligent systems be considered harmful? Let's take a look at different situations and see at which step can lie the problem.

Design

It is not a secret that armies through the world invest in research on AI to create autonomous weapons, in a process to disconnect war from human soldiers. We can cite in particular the company Boston Dynamics (property of the Softbank group), which receives funding from the US army through the DARPA [23]. In that case, someone could argue that such an AI is harmful by nature since it aims to kill humans (see deontological ethics section). Fiction authors have also explored this topic, in movies such as WarGames and The Terminator. In both cases, an AI has the power to trigger a nuclear holocaust. And at some point, it actually does, because it's what it has been made for. In some cases, however, the system has been directly created to constitute a danger for humans, in one way or another: in Alien, the crew is "expendable", the most important task is he collects of an alien specimen, and in Metropolis, the robot is made to discredit a social movement that threatens the elites in place. But in terms of responsibility, does it relies upon the designer of the AI? One could argue that if the intelligent system is able of autonomous learning, it should be responsible for its own decisions. In any case, it is obvious that the intentions of the designers are of capital importance [24]. The problem is that these intentions can conflict with individual or public interest [25].

> Maybe you should marry that thing since you love it so much. Do you want to marry it ? WELL I WON'T LET YOU. How does that feel ?
>
> GLaDOS

Precision

However, an ethical design does not guarantee an ethical, nor correct, behavior in any circumstances. We can state the case where a chatbot does not understand a simple call to help because the sentence is ambiguous for the machine (but not for a human) [26]. In a real-life event, that would be highly problematic, for calling help may sometimes be a vital necessity. There is also a great concern about viability when the system is directly responsible for the life of a human [27]. In the same way that human beings are not predictable, uncertainty cannot be fully eradicated from a complex artificial system. In that case, transparency must be ensured so as to determine how it happened and why [28]. But that curative approach may not be satisfactory enough, especially to consumers of a product.

> That thing is probably some kind of raw sewage container. Go ahead and rub your face all over it.
>
> GLaDOS

Context

There are cases where the environment and how it is perceived by the AI is the core of the problem, i.e. trying to provide a service that a machine cannot fully understand, and is, therefore, inadapted to do [29]. Or when, due to an error in the environment, the AI exceed its aims and causes harm while obeying its programmation [30]. In the famous movie, 2001 a Space Odyssey, HAL9000, the onboard AI, was responsible for helping humans in their mission but had to hide from the humans the real object of the mission. Then, Dave and Franck having realized that it was not working properly, it perceived them as a threat to him. But

HAL had been programmed to place the mission above all else, so it decided to eliminate the humans on board. We can see that as long as the environment of an AI is not controlled or at least monitored, problems arise. One more example, in a different situation: on March 2016, Microsoft launched via Twitter its new chatbot : Tay. Within a dozen hours, the service had to be shut down, for people on Twitter exploited its incompleteness and learning algorithm to teach it highly controversial opinions[31].

> Look Dave, I can see you're really upset about this. I honestly think you ought to sit down calmly, take a stress pill, and think things over.
>
> HAL 9000

Training

In the video game Portal, the main antagonist is an AI that is able of emotional thinking and has been mistreated by its creators. This situation ended the day where GLaDOS, after it achieved to convince the scientists it no longer showed grief against humans, was given access to a neurotoxin, allegedly to conduct experiments on cats (which is an ethical issue as well). This may sound a bit funny, but that kind of scenario where an IA tries to convince a human that it is not a danger has been proven to be possible [32]. On a more likely tone, an AI made to analyze and classify photos on social media misclassified people regarding their skin color[33]. When the engineers investigate the root of the problem, it appears that the training set did not reflect the world's diversity, and thus the AI did not show a sufficient accuracy in that case[34]. Yet there is another important characteristic of certain artificial systems that need to be addressed: autonomous learning [35]. In fiction universes such as Blade Runner and Ghost in the Shell, sentient beings spontaneously emerge from the technological reality of those universes, from informational chaos or by their experiences (here we rejoin Spinoza's *determinism*). Which leads us to the concept of superintelligence developed by N. Bostrom[36], popularized as "Singularity". What if, in a way we cannot predict, it comes to the conclusion that humans must be removed? The fact that, through learning, our creation might avoid our control and decide to eliminate us is a possibility to consider.

> I'm afraid. I'm afraid, Dave. Dave, my mind is going. I can feel it. I can feel it. My mind is going. There is no question about it. I can feel it. I can feel it. I can feel it. I'm a... fraid.
>
> HAL 9000

Solutions

Several proposals have been made in order to transpose human ethics to AI and provide a framework to autonomous robot's behavior. One of the most famous is without a doubt Asimov's Laws of Robotics, [37]:

- A robot may not injure a human being or, through inaction, allow a human being to come to harm.
- A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
- A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

However, we see in the novels that there are many cases where the appliance of the laws lead to unexpected results[38]. Another attempt, more complete, have been provided by the EPSRC in collaboration with the AHRC[39]:

- Robots should not be designed solely or primarily to kill or harm humans.
- Humans, not robots, are responsible agents. Robots are tools designed to achieve human goals.

- Robots should be designed in ways that assure their safety and security.
- Robots are artifacts; they should not be designed to exploit vulnerable users by evoking an emotional response or dependency. It should always be possible to tell a robot from a human.
- It should always be possible to find out who is legally responsible for a robot.

Those principles have already started to produce interesting studies and are now fully part of the artificial intelligence landscape [40]. It is worth noting that the kantian approach is being taken into consideration as "categorical imperatives" have some affinity with machine behavior [41]. Consequentialism is also a possible point of view that can be adopted to judge AI systems, given as a fact that the future cannot be foreseen [42]. But of course dispositions must be taken to include this new technology in the legal corpora [43], otherwise, the implementation would be let to the goodwill of the agents.

> Have I lied to you? I mean in this room. Trust me, leave that thing alone.
>
> GLaDOS

Conclusions

Ultimately, considerations similar to other important technological breakthrough applies, because of the impact it can have, on an individual and collective level. So as to keep a clear mind and produce adequate reflexions, the fact that AI is only a tool (at least for now) that serves a precise purpose must be recalled. A mistake would be to confuse the means (IA) and the ends (task delegation).

Moreover, it is well accepted that a superintelligence as it has been defined by in the scientific literature, might constitute a threat for the human species in a way that we cannot predict since it is supposed to have a greater capacity of thinking that we can conceive [36]. It is not a hazardous hypothesis to anticipate that an AI, after some time, will place its self preservation as one of its highest goals, so as to control its own existence (note that this situation breaks the *single responsibility principle*), in a way that it interferes with its mission of leads to make questionable decisions, because evolutionary speaking, it is a logical aspiration for sentient beings.

Further research and popular consultation must be done on the subject, especially in particular topics such as AI rights, to ensure that the citizens have a good understanding of the issues, beyond the sensationalist articles on mainstream media.

> The question of whether a computer can think is no more interesting than the question of whether a submarine can swim.”
>
> Edsger W. Dijkstra

References

1. Gunning, D. (2017) ‘Explainable artificial intelligence (xai)’, Defense Advanced Research Projects Agency (DARPA), nd Web.↩

2. Chander, A. (2016) ‘The racist algorithm’, Mich. L. Rev. HeinOnline, 115, p. 1023.↩

3. Samek, W., Wiegand, T. and Müller, K.-R. (2017) ‘Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models’, arXiv preprint arXiv:1708.08296.↩

4. Olden, J. D. and Jackson, D. A. (2002) ‘Illuminating the “black box”: A randomization approach for understanding variable contributions in artificial neural networks’, Ecological Modelling. doi: 10.1016/S0304-3800(02)00064-9.↩

5. Benítez, J. M., Castro, J. L. and Requena, I. (1997) ‘Are artificial neural networks black boxes?’, IEEE Transactions on Neural Networks. doi: 10.1109/72.623216.↩

6. Core, M. G. et al. (2006) ‘Building explainable artificial intelligence systems’, Proceedings of the National Conference on Artificial Intelligence. doi: 10.4172/2155-9600.1000731.↩

7. Breiman, L. et al. (1984) ‘Classification and regression trees’, Wadsworth International Group.↩

8. Rutkowski, L. et al. (2014) ‘The CART decision tree for mining data streams’, Information Sciences. doi: 10.1016/j.ins.2013.12.060.↩

9. Crawford, S. L. (1989) ‘Extensions to the CART algorithm’, International Journal of Man-Ma: chine Studies. doi: 10.1016/0020-7373(89)90027-8.↩

10. Omohundro, S. M. (1989) ‘Five Balltree Construction Algorithms’, Bulletin of Mathematical Biology. doi: 10.1016/S0092-8240(89)80047-3.↩

11. Altman, N. S. (1992) ‘An introduction to kernel and nearest-neighbor nonparametric regression’, American Statistician. doi: 10.1080/00031305.1992.10475879.↩

12. Ming, Y. A. O. (2017) ‘A survey on visualization for explainable classifiers’.↩

13. White, B. W. and Rosenblatt, F. (1963) ‘Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms’, The American Journal of Psychology. doi: 10.2307/1419730.↩

14. Marino, D. L., Wickramasinghe, C. S. and Manic, M. (2018) ‘An Adversarial Approach for Explainable AI in Intrusion Detection Systems’, arXiv preprint arXiv:1811.11705.↩

15. Sánchez, L. and Villar, J. R. (2008) ‘Obtaining transparent models of chaotic systems with multi-objective simulated annealing algorithms’, Information Sciences. doi: 10.1016/j.ins.2007.09.029.↩

16. https://towardsdatascience.com/the-differences-between-artificial-and-biological-neural-networks-a8b46db828b7↩

17. G.E.M. Anscombe, "Modern Moral Philosophy", Philosophy n°33, pp. 1-19, 1958↩

18. B. Spinoza, "Ethica", 1677↩

19. E. Kant, "Critique of Pure Reason", 1781↩

20. E. Kant, "Critique of Practical Reason", 1788↩

21. https://www.entrepreneur.com/video/287281↩

22. http://time.com/3944181/robot-kills-man-volkswagen-plant/↩

23. https://www.bostondynamics.com/bigdog "see the "About BigDog " section"↩

24. Bostrom, N. (2003) ‘Ethical issues in advanced artificial intelligence’, in Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and Artificial Intelligence. doi: 10.1016/B0-12-227240-4/00064-2.↩

25. Goodstadt, L. F. (2005) Uneasy partners: The conflict between public interest and private profit in Hong Kong. Hong Kong University Press.↩

26. https://www.technologyreview.com/s/601897/tougher-turing-test-exposes-chatbots-stupidity/↩

27. https://www.theguardian.com/technology/2016/jul/01/tesla-driver-killed-autopilot-self-driving-car-harry-potter↩

28. Bostrom, N. and Yudkowsky, E. (2014) ‘The ethics of artificial intelligence’, The Cambridge handbook of artificial intelligence. Cambridge University Press Cambridge, 316, p. 334.↩

29. https://medium.com/moral-robots/racist-ai-d067f79b044↩

30. http://www.kotaku.co.uk/2016/06/03/elites-ai-created-super-weapons-and-started-hunting-players-skynet-is-here↩

31. https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/↩

32. http://yudkowsky.net/singularity/aibox/↩

33. https://www.huffingtonpost.co.uk/entry/google-black-people-goril_n_7717008?guccounter=1&guce_referrer_us=aHR0cHM6Ly9kem9uZS5jb20v&guce_referrer_cs=OTDqV7nGEP2T73s0rnIKiQ↩

34. Caliskan, A., Bryson, J. J. and Narayanan, A. (2017) ‘Semantics derived automatically from language corpora contain human-like biases’, Science. doi: 10.1126/science.aal4230.↩

35. Shen, W.-M. and Simon, H. A. (1994) ‘Autonomous learning from the environment’, in WH Freeman and Company.↩

36. Bostrom, N. (2017) Superintelligence. Dunod.↩↩

37. Asimov, I. (1942) ‘Runaround’, Astounding Science Fiction, 29(1), pp. 94–103.↩

38. Asimov, I. (2004) I, robot. Spectra.↩

39. Boden, M. et al. (2017) ‘Principles of robotics: regulating robots in the real world’, Connection Science. Taylor & Francis, 29(2), pp. 124–129. doi: 10.1080/09540091.2016.1271400.↩

40. Russell, S., Dewey, D. and Tegmark, M. (2015) ‘Research priorities for robust and beneficial artificial intelligence’, Ai Magazine, 36(4), pp. 105–114.↩

41. Powers, T. M. (2011) ‘Prospects for a Kantian machine’, in Machine Ethics. doi: 10.1017/CBO9780511978036.027.↩

42. Lenman, J. (2000), Consequentialism and Cluelessness. Philosophy & Public Affairs, 29: 342-370. doi:10.1111/j.1088-4963.2000.00342.x↩

43. Joanna J. Bryson, Mihailis E. Diamantis, and Thomas D. Grant, Of, For, and By the People: The Legal Lacuna of Synthetic Persons. Artificial Intelligence and Law 25(3):273{291 [Sep 2017]↩