Inherent Limits of the Use of AI (Deep Learning) in Health Care

Jörn Erbguth
Institute of Information Service Science
University of Geneva, Switzerland
joern@erbguth.net

***Abstract*—AI is being increasingly used in health care. Most concerns are about possible biases of such AI systems and possible ways to counter these biases are being discussed. However, bias is rather a symptom of a much larger issue and introducing counter bias threatens to render things worse. The proposed EU AI regulation is taking the right regulatory approach but has a much too broad definition of AI.**

Keywords—AI, Deep Learning, bias, health care, AI regulation

I. Introduction

Currently the most used AI technique is Deep Learning. Deep Learning is a type of Machine Learning. It is different from conventional computing where a program implements an algorithm to solve a problem. Deep Learning is based on artificial neural nets. These show several interesting and several problematic properties. It is important to understand these characteristics to recognize what is behind, for example, TESLA's autopilot failures or bias in AI systems. When using these AI systems, people must be aware of these limitations and only use this technology in a way in which the effects of these inherent problems can be tolerated. This paper will also address how the proposed EU regulation tries to address these problems.

II. Machine Learning

Machine Learning is part of AI and is essentially the automated detection of patterns in data [1]. These patterns can then be used to simulate intelligent decision-making in examples that were not present in the training dataset. While storing data only reproduces the original data, learning can abstract from or interpolate training data, and can even work with partly contradicting or incomplete training datasets. Machine Learning is neither programmed nor does it implement explicit rules. It simply detects some patterns in the data without questioning their origin nor whether they can be generalized as rules.

III. Deep Learning

Deep Learning is one among many different machine learning techniques. It uses artificial neural networks. This is a system design inspired by the way the biological nervous systems and the brain work and is part of *soft computing* techniques. These techniques deal with partial truth, uncertainty, and approximation to solve complex problems [2]. There are a variety of architectures of the artificial neural network and a variety of ways to train it. However, training an artificial neural network always requires a large dataset of training data, a random initial state and an optimization algorithm that modifies the weights in the network with each training cycle.

Artificial Neural Nets were studied as early as in 1948 by von Neumann [3]. He compared what he knew at that time about the human brain with the type of computers of that time that were built from vacuum tubes. From his reasoning about the human brain, he designed an artificial neural net that would work differently to conventional computers, be resilient to errors and would not need any programming. However, due to lack of computing power at this time, his approach was not widely used for a long time.

IV. Similarity to the Human Brain

Most times people do not learn by rules or logic but by examples and by feedback. We can abstract from examples without even consciously knowing the resulting ruleset. Being subconsciously trained with many examples, we can recognize patterns in incomplete data or even inconsistent datasets. Doctors and lawyers are often confronted with incomplete datasets. Their experience helps them to still arrive at a good diagnosis or a legal analysis. Just by looking at a case, experts have a feeling about the right solution. This type of speedy analysis when only incomplete data is available can be very useful. For example, it might have been useful in the past to detect a threat and run away before complete data is available for a thorough analysis.

However, this intuitive experience-based system also has its failures. First, we do not know how well the brain interpolates our knowledge for unknown situations. A little child that has never seen a cow, but only dogs might call the first cow she sees a dog. The little child might also call it a horse. There is some probability to both or some other alternative, but we simply do not know what will happen.

Second the brain might mistake some correlation as a rule. A person that has always seen male CEOs accompanied by female secretaries will automatically deduct that the female he encounters at a business meeting is the secretary and not the CEO.

Humans also have the possibility to consciously deal with knowledge. When confronted with the CEO and the secretary, they know that their experience might lead them to false assumptions and counteract their otherwise existing gender stereotypes. Doctors and lawyers undergo long training sessions where they learn legal and medical analysis which questions their intuitive result. The analysis either confirms or refutes their first assumption.

Interestingly, mainly in the 1980s but also later, different types of AI systems called *expert systems* were designed. They were not self-learning but required very skilled knowledge experts. These knowledge experts interviewed domain experts about their knowledge, broke it into very small logical pieces and an inference engine used these knowledge pieces to answer questions. These expert systems were the exact opposite of current artificial neural nets. At every point they adhered to clearly stated rules. Nothing was a black box. An example is a heart failure monitoring system[4]. However, these systems can only cope with very specific situations and are challenged by complex realities. Much of the knowledge of domain experts was not conscious knowledge and when confronted with the real world, rules tended to become overly complex, still not accurate enough and difficult to maintain.

V. Systematic Strengths and Failures of Deep Learning

One strength of Deep Learning is the ability to "learn" from incomplete and even inconsistent datasets. Instead of requiring knowledge engineers, who try to find out why

human experts decide the way they do, Deep Learning can learn directly from data. While expert systems would break if their ruleset contained a single contradiction, Deep Learning can be trained fairly well with bad data. It needs neither any domain knowledge nor rules.

The training of Deep Learning system varies. It basically consists of a random initial state of the network, a set of training data and a set of validation data. Starting from the random state, the system is trained using the training data. *Trained* means that the network processes the input data of the training data and then the resulting output is compared with the desired output to calculate the error. The difference is propagated back to the individual neurons and the weights of the connections between the neurons are modified so that the error be reduced.

After a while the error cannot be lowered any more. The validation dataset is then fed into the network. If the error is sufficiently low, the trained network is selected, if the error is too high, a new initial state is taken and trained with the same data.

Many factors can influence the quality of the training result: these include the number of neurons, the topology and connectivity of the network, the training algorithm, the learning speed, the random initial state of the neural network, etc.

When the validation dataset produces good enough results, a final test dataset is often used to verify the resulting system. Every test dataset only evaluates the system's performance for a very limited set of data points. Deep Learning is based on the hope that the network has generalized well the datasets it saw during the training phase. Since many possible systems might have been validated with the validation dataset, a good result might be accidental. Therefore, the final test is not performed with the validation dataset, but with a different test dataset that has not been used before.

This procedure results in several weaknesses of a trained Deep Learning system:

- We can only hope that the system will handle cases outside the datasets used in the training reasonably well, since we performed some tests. However, we cannot guarantee any minimal quality. The risk of failure in areas with little or no training data is higher, but even in areas with many data points, completely wrong results are possible.
- We can almost be sure that the system will mistake some false correlations as rules. This is the main cause for bias or perceived discrimination of Deep Learning systems.
- Due to the initial random state, every training session run with the same data will result in a different system.
- The system is a black box. We could analyze the system to find out what the system's learned rules are. Often, the result is quite complex and full of disturbing noise. Analyzing the original training data should give far better results and might allow a conventional, rules-based system to be built.
- In some situations, parts of the training data might be reconstructed from the trained network. This is a serious problem if, for example, personal health data is used for the training.
- Finetuning the input values to the system can be used to manipulate the system [5]. Due to the characteristics of Deep Learning, there are many input values that lead to false results. Finding those that are close to a given input value can help to arbitrarily manipulate these systems. This is also called *hacking* of Deep Learning systems.
- The training dataset has an influence on the training result. A biased training set might lead to a biased trained system. However, a perfect training set will also lead to a biased trained system due to it picking up false correlations. There is a big influence of the training dataset on the quality of the trained system. However, due to the error tolerance of the training procedure, a small bias in the training data should not be the main source of bias in the trained system.
- Training datasets could include special values that act as invisible backdoors in the trained system.

VI. Application of Deep Learning in the Health Sector

Given all these disadvantages, should Deep Learning be used at all in the health sector? Let's look at an example: X-ray diagnosis.

X-ray images are used to detect pneumonia [6], COVID19 [7] or lung cancer [8], for example. Deep Learning has also been tested to detect breast cancer in mammograms [9]. The authors of these papers report good performances of these systems. An area of great concern is bias in the training, validation and test datasets. *Seyyed-Kalantari et al* evaluated bias in trained systems [10]. They trained neural nets with annotated X-ray images. They discovered that while the overall quality was satisfactory, the quality seemed to vary depending on the respective population (age, sex, "race" and insurance status). The researchers discovered that the group with social disparities had the least favorable results. The difference in quality might be explained by two groups of reasons:

a) The system has a different recognition quality for different population groups.

b) The test data has a different amount of wrongly classified images depending on the population group.

Regarding option a) it seems rather unlikely that a system can detect the skin color or insurance status of a person by analyzing an X-ray. However, *Banerjee et all* [11] claim that they successfully trained a Deep Learning system to detect the skin color of a person by analyzing one of their X-ray images. This work has not been peer-reviewed yet. The issue with Deep Learning is that it is impossible to determine whether a system can identify the skin color or some other undetected property that happens to be correlated in the test set. Even if the direct recognition can be excluded, specific conditions might occur with different frequencies in different population groups. For example, the diagnosis of X-rays of obese people is more difficult and error-prone than of normal-weight people [12]. If one population group has a different percentage of obese people, the average diagnosis quality of this group will be lower. It can be questioned whether this would constitute a bias in the system. When task difficulty is not equal, the results will not be equal either.

While correct X-ray diagnoses of X-rays from obese patients are more difficult for humans and probably also for artificial neural nets, other deficiencies could be due to insufficient training data. If cases backed by insufficient training data occur more frequently in some population groups, these population groups seem to receive a lower quality of automated diagnoses.

Some errors in the test dataset might be more present in the test cases of one population group. Errors in test datasets indicate errors where the system is offering correct results. If these errors are distributed differently in different population groups, it will indicate a bias that does not exist (option b).

Finally, since the training of a neural net is not a deterministic process, errors are always introduced randomly. If retraining the system multiple times shows the same bias, the random initial state can be excluded as an explanation of the bias.

Regarding option b) wrongly classified test data will lead to a consistent bias in the test result but does not correspond to a real bias of the system. *Seyyed-Kalantari et al* fail to even consider that their detected bias is a bias in their testing.

VII. Possible remedies

While a bias in the task difficulty can hardly be remedied and a bias in the quality of the test dataset does not have an impact on the system quality, it makes perfect sense to consider that even relatively rare conditions are included with a sufficient number of training images. Several techniques are discussed to reduce bias [13]. However, those techniques have a high probability to render things worse.

One example is post processing to counter some gender bias that has been detected. A voluntary counter bias is added to the result [14]. Adding counter bias will result in treating identical cases differently based on factors like race or sex. The perceived bias that should be countered depends on the test dataset. Test datasets are biased as are training datasets. Even worse, test datasets can be tuned to use training artefacts of a neural net to indicate any amount of bias, without looking suspicious. Therefore, counter bias does not remove bias but adds additional (but different) bias to a system based on a measurement that is itself exposed to obvious and hidden bias.

Currently, this is often prohibited without consent by Art. 9 GDPR in the EU but UK plans to change UK-GDPR to allow *bias correction* [15]. This has a high risk to result in direct discrimination based on protected attributes and raises serious human rights issues. Concerning the analysis of X-rays, adding post processing counter bias to decrease the quality of cancer detection for people outside protected population groups is fortunately not an option proposed.

VIII. Practical application of these systems

There are three reasons for using these systems:

- They provide better quality diagnosis.
- They can provide diagnosis of a larger population more frequently since medical personnel is a limited resource.
- They are cheaper.

There is a high potential that the use of such systems will have a positive health impact. However, there is also a risk that these systems will be merely used for cost reduction even when the quality does not reach the quality of a human doctor.

IX. Risks

While a lot of public attention focuses on the risk of bias towards specific population groups, bias is neither the only nor the biggest risk involved. The biggest risk is the unpredictability of the trained systems. We do not know whether a good test result will replicate in practical use of a system. Some small modification of the images presented might already lead to completely different results. This has been extensively shown for image recognition [16]. Self-driving cars are having accidents because of recognition errors [17]. This creates a high risk when a system is used for early detection of illnesses without parallel medical supervision. Therefore, it is important to accompany any use of these systems with constant supervision by medical experts who can review a substantial percentage of cases to discover any anomalies at a very early stage. Testing a Deep Learning system does not provide enough certainty since it is not known which deviation from the test dataset might lead to erroneous results.

X. Regulatory Approach

The European Union discusses a new regulation on AI [18]. It takes a risk-based approach and categorizes applications into three different risk categories:

- Unacceptable risk
- High risk
- Low or minimal risk

While applications that create an unacceptable risk shall be prohibited, high risk applications shall be heavily regulated. High-risk applications include applications that are already governed by safety regulation like medical devices. High-risk applications also include:

- Biometric identification
- Management of critical infrastructure
- Education and vocational training
- Employment
- Access to essential public and private services
- Law enforcement
- Migration, asylum and border management
- Justice and democratic processes.

The proposed regulation mandates for high-risk applications:

a) Risk management throughout the entire life cycle (Art. 9)

b) Training, testing and validation data governance (Art. 10)

c) Technical documentation (Art. 11)

d) Record keeping (Art. 12)

e) Transparency and information of users (Art. 13)

f) Human oversight (Art. 14)

g) Accuracy, robustness and cybersecurity (Art. 15).

It is yet to be seen whether Deep Learning systems will be able to fulfill all requirements for high-risk systems, particularly regarding transparency and sufficient reliability. Human oversight is one of the central requirements.

The proposed regulation is limited to AI. However, the definition of AI in appendix I is so broad that it virtually includes all existing hard- and software. Only some articles provide exceptions, e.g., Art. 10 applies only to systems that are trained. Therefore, a simple X-ray system that allows a patient's name to be queried might be subject to the same regulation as a sophisticated Deep Learning system.

XI. Conclusion

While public attention focuses on bias and discrimination, the risks of applying AI (Deep Learning) in the health sector are far broader. The slightest modification of input data might lead to a complete malfunction of a system. Constant supervision is mandatory.

[1] S. Shalev-Shwartz and S. Ben-David, Understanding Machine Learning: From Theory to Algorithms. Cambridge: Cambridge University Press, 2014. doi: 10.1017/CBO9781107298019.

[2] D. Ibrahim, 'An Overview of Soft Computing', Procedia Comput. Sci., vol. 102, pp. 34–38, 2016, doi: 10.1016/j.procs.2016.09.366.

[3] J. V. Neumann, Theory of Self-Reproducing Automata, First Edition. University of Illinois Press, 1966.

[4] E. Seto, K. J. Leonard, J. A. Cafazzo, J. Barnsley, C. Masino, and H. J. Ross, 'Developing healthcare rule-based expert systems: Case study of a heart failure telemonitoring system', Int. J. Med. Inf., vol. 81, no. 8, pp. 556–565, Aug. 2012, doi: 10.1016/j.ijmedinf.2012.03.001.

[5] W. Brendel, J. Rauber, and M. Bethge, 'Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models', ArXiv171204248 Cs Stat, Feb. 2018, Accessed: Oct. 16, 2020. [Online]. Available: http://arxiv.org/abs/1712.04248

[6] E. Ayan and H. M. Unver, 'Diagnosis of Pneumonia from Chest X-Ray Images Using Deep Learning', in 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT), Istanbul, Turkey, Apr. 2019, pp. 1–5. doi: 10.1109/EBBT.2019.8741582.

[7] S. R. Nayak, D. R. Nayak, U. Sinha, V. Arora, and R. B. Pachori, 'Application of Deep Learning techniques for detection of COVID-19 cases using chest X-ray images: A comprehensive study', Biomed. Signal Process. Control, vol. 64, p. 102365, Feb. 2021, doi: 10.1016/j.bspc.2020.102365.

[8] W. Ausawalaithong, A. Thirach, S. Marukatat, and T. Wilaiprasitporn, 'Automatic Lung Cancer Prediction from Chest X-ray Images Using the Deep Learning Approach', in 2018 11th Biomedical Engineering International Conference (BMEiCON), Nov. 2018, pp. 1–5. doi: 10.1109/BMEiCON.2018.8609997.

[9] X. Zhu, J. Shi, and C. Lu, 'Cloud Health Resource Sharing Based on Consensus-Oriented Blockchain Technology: Case Study on a Breast Tumor Diagnosis Service', J. Med. Internet Res., vol. 21, no. 7, p. e13767, 2019, doi: 10/gf5jgg.

[10] 'CheXclusion: Fairness gaps in deep chest X-ray classifiers'. https://www.worldscientific.com/doi/epdf/10.1142/9789811232701_0022 (accessed Sep. 07, 2021).

[11] I. Banerjee, A. Bhimireddy, J. Burns, L. Celi, and L. Chen, 'Reading Race: AI Recognizes Patient's Racial Identity In Medical Images', p. 46.

[12] 'An X-Ray Tech's Guide to Radiography in Obese Patients', CE4RT, Oct. 30, 2013. https://ce4rt.com/rad-tech-talk/radiography-and-obese-patients/ (accessed Sep. 07, 2021).

[13] T. Feldman and A. Peake, 'End-To-End Bias Mitigation: Removing Gender Bias in Deep Learning', ArXiv210402532 Cs, Jun. 2021, Accessed: Sep. 11, 2021. [Online]. Available: http://arxiv.org/abs/2104.02532

[14] M. Hardt, E. Price, and N. Srebro, 'Equality of Opportunity in Supervised Learning', ArXiv161002413 Cs, Oct. 2016, Accessed: Sep. 11, 2021. [Online]. Available: http://arxiv.org/abs/1610.02413

[15] P. Driscoll, 'Data: a new direction', p. 146.

[16] D. Heaven, 'Why deep-learning AIs are so easy to fool', Nature, vol. 574, no. 7777, pp. 163–166, Oct. 2019, doi: 10.1038/d41586-019-03013-5.

[17] C. I. and P. V.-D. Business CNN, 'Tesla is under investigation because its cars keep hitting emergency vehicles', CNN. https://www.cnn.com/2021/08/16/business/tesla-autopilot-federal-safety-probe/index.html (accessed Sep. 07, 2021).

[18] 'EUR-Lex - 52021PC0206 - EN - EUR-Lex'. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206 (accessed Sep. 07, 2021).