Ethical Artificial Intelligence

Yiqiao Yin
W.Y.N. Associates, LLC

Abstract

This assignment investigates the ethical problems in Artificial Intelligence or Ethical AI. We explore the controversial area of this branch of research and explains the reasoning behind where the controversial areas come from. In addition, the report surveys a list of papers about potential solutions to address Ethical AI. This report lands on future research avenues that could potential resolve the controversies of discussed in this field.

1 Ethical Artificial Intelligence (EAI)

There is an exploding number of fields of applications in adopting Artificial Intelligence or AI based solutions in business practice specifically when it comes to replacing part of the decision-making process such as healthcare, financial risk analysis, high-frequency trading, autonomous driving, and so on. Many news have been accusing autopilot of Tesla [1] Take autonomous driving as an example, we can denote two simple classes of actions: "drive" or "stop". In extreme situation such as seeing a pedestrian on the highway, the decision "drive" may fatally harm the pedestrian while the decision "stop" may fatally harm the driver and other drivers behind them. While claims can be made easily, the solution of fixing this problem is more sophisticated than what the news has surfaced. In statistical machine learning, a decision making process can be summarized in a confusion matrix, which the two classes of actions can be used. An adaption of such confusion matrix is drawn in Figure 1. The decision of "drive" or "stop" cause two classes of actions. The condition can be either predicted from AI algorithms in autopilot or the actual conditions in reality (what the autopilot should have done). If "drive" is safer in actual condition and the autopilot does exactly that, there is no casualty. If "stop" is safer in actual condition and the autopilot does exactly that, there is no casualty either. These two scenarios are called true positives and true negatives, respectively. In other words, no one is at danger under true positive and true negative situation. If the actual condition should have been "drive" yet the autopilot predicts "stop", the pedestrian is probably okay yet this poses a great danger for the driver to crash, which can also harm the nearby vehicles. This is called false negatives and it is also known as Type II error. If the actual condition should have been "dtop" yet the autopilot predicts "drive", the pedestrian is probably going to be in grave danger but the driver and nearby vehicles are probably okay at that moment. This is called false positives and it is also known as

[1] Source: https://www.theverge.com/2016/6/30/12072408/tesla-autopilot-car-crash-death-autonomous-model-s.

Type I error. There is an innate trade-off between Type I and Type II errors, which is why most AI-based solutions turned out to be an ethical problem not a machine learning problem.

		Predicted Conditions	
		Drive	Stop
Actual Conditions	Drive	True positives	False negatives (II)
	Stop	False positives (I)	True negatives

Figure 1: **Confusion matrix as decision-making process**. The Type II error implies that the driver may get hurt. The Type I error implies that the pedestrian may get hurt.

In this specific example illustrated in Figure 1, another way of reviewing the ethical problem is using the "trolley problem" (which can be demonstrated in Figure 2.

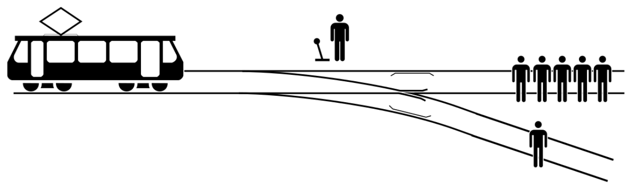

Figure 2: **Trolley problem**

Due to this analogy, many interests have arise in investigation of the ethical problems in AI especially focusing on autonomous agents Lin et al. (2014); Yampolskiy (2013); McLaren (2006); Moor (2006); Russell et al. (2015).

2 Moral Philosophy

The first section raised the topic of the potential moral philosophy involved in machine learning and AI. The moral rules are taught to us in early adulthood and these rules can be related to some principles or religions. We grow up following these doctrines and they allow us to put a tag on behaviors where we can classify "good" from "evil". Disregard the background and personal history, certain moral rules can be normative to practice across all different race of people and religions. These moral rules form principle of ethics. As such, there are three important approaches when it comes to practicing ethics in the literature Cointe et al. (2016); Christman and Zalta (2015). The first one is called "virtue ethics" which states that a person is ethical if and only if he acts according to the values from certain doctrine. The second one is called "deontological ethics" which states that a person is ethical if and only if he respects the principle obligations related to possible situation. The third one is called "consequentialist ethics" which states that a person is ethical if and only if he rationally evaluates the moral consequences and chooses the one that has the most moral values.

The controversy arises under the following scenarios: (1) The machine is used to replace human completely in decision-making process. A good example can be a lazy driver having complete faith for autonomous driving and decides to take a nap while the vehicle is in motion. However, the fact that this is even allowed at the first place also deserve attention and should be discussed. Disregard the responsible party involved if an incidence happens, the first scenario would raise a series of questions where and how ethical values are executed. (2) The machine is used without questioning the data. In this scenario, the machine is considered to be well trained and is able to generate high performance. Before the machine is put in production, there should be a protocol to evaluate the data source and search for potential bias in the data. If the data is biased, no matter how accurate the machine is it will create biased results. A good example can be criminal database profiling and predicting using "biased" data [2] A debate rises whether to include gender and racial information for the criminals. On one hand, the data is collected with visual appearance enforced with police force on the field. Hence, the data is collected with the prior knowledge of the police at the location real time. The moral principles, disregard correct or not, is injected in the data collection process. After modeling and machine learning, the predictions even from well trained models will learn at best the bias created in the labels of the training data. If these features are removed, another debate arises why data manipulation is at presence and the consequence related to data manipulation. (3) The machine is too complicated to understand and the decision making process generates a system without the technical expertise. A well known example is in financial risk assessment. A private project I have seen before is in mortgage rate prediction and analysis. An American citizen could apply for a mortgage rate with a list of his information. A machine is then deployed to analyze the default rate based on the client's information. This can happen because the banks need to assess the level of risk present with this client, which is then used to understand what rate to provide to this client. The machine algorithms can be as simple as linear regression or as complicated as a deep neural network. However, the better the performance the more complicated the machine is. Hence, it is not exactly transparent to the management team who are making the decisions and who do not have the technical background. This pose a grave challenge, because the machine is considered as a black-box to the decision maker. While each machine is built with premises and decision rules, these assumptions are not always clear to the decision makers or whoever is utilizing their outcome.

3 Ethical Judgement Process by Cointe et al. (2016)

The work by Cointe et al. (2016) proposed an ethical judgement process or a goodness process to evaluate the morale fit of a system. A decision making process using a machine learning algorithm can be designed into a system of which the authors propose to evaluate using the goodness model.

To conduct this evaluation procedure, the authors proposed two requirements. First, a list of threshold is discussed and proposed to form what are consisted of moral values. In other words, the finalized moral value is a function of different moral key words such as awareness, goodness, ontology, belief, generosity, honesty, and so on. Second, the process of ethical judgement is proposed to be on a various of scale of good and right. This means that certain evaluation is numerically computed and some scoring system is at place.

The main diagram of the proposed model is presented in Figure 3. Based on the mental states and independent of particular architecture, a global representative view of the proposed goodness model can be presented. The arrows in the diagram of Figure 3 refers to data flow. The awareness process is marked with the grey box. The evaluation process is marked with boxes that are less grey than the awareness process. There are also

[2] NYPD Criminal Database News (https://www.techtarget.com/searchbusinessanalytics/news/252459511/NYPDs-Patternizr-crime-analysis-tool-raises-AI-bias-concerns).

goodness process and it is followed with the rightness process. The goodness process are the least grey and the rightness process is the white box. In addition, the diagram also contributes knowledge base in evaluation process and the rightness process. In this case, an ethical judgement process or EJP follows the definition that is a function of Awareness Process or AP, Evaluation Process or EP, Goodness Process or GP, and Rightness Process or RP.

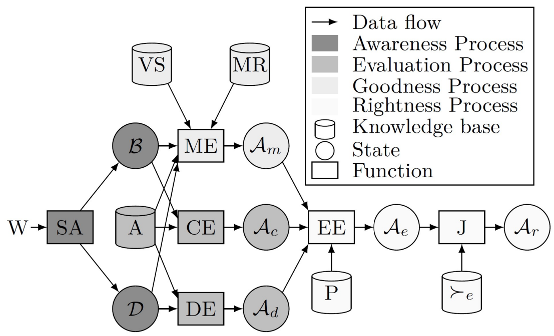

Figure 3: **Global Model**.

In addition to understand the moral principles, the next step is the capability to execute the moral principles. This perspective is vaguely mentioned in the authors' work, because execution is the next important component for the system to be enforced. Without execution, the system is empty and can be enforced. In other words, protocols need to be put in place shall certain orders not be executed. In the NYPD bias case Badr and Sharma (2022), there is a trade-off between how much data is collected and what data is allowed to be collected. From machine learning perspective, it is important to collect as much data as possible. From the perspective of the crowd, no one wants to release any data. In addition, discussions also raise awareness of what type of data to collect. From ethical point of view, the gender and race are removed from the data.

The execution also brings up another important issue. Since the algorithm and the system are constructed by humans and they are just the logical extension of human will, it is then essential to discuss whose will is in the system. Are we using teachers, researchers, priests, or some other intelligent people telling us what constitutes good moral ethical ground? This part of the debate is not thoroughly discussed in the paper either. On the other hand, it is not quantifiable which could be why the authors' have left this part out. If we agree that the principles of moral values are created by a team of scholars and well respected individuals, then the execution and enforcement can easily turn the system into a dictatorship, which is by itself contradictory to the purpose of why we have this moral evaluation system set up in the beginning.

4 Explainability and Interpretability

One part the authors left out was the explainability and interpretability of a machine. Though it is easy to mention it, it is difficult to discuss in depth and obtain an agreed upon definition of explainability and interpretability of a machine or an algorithm.

The level of understanding to its human user or end user can be commonly considered as explainability or interpretability. Depending on where the source of the information originate from, there can be explainable machines and interpretable machines. Like the linguistics concept of external and internal, the explainability refers to external effort and process to make the models understandable to its end-users while the interpretability refers to the internal capability to make inference of its results.

The classical example of an interpretable model is linear regression. Given an independent variable X and a dependent variable Y, a linear model $Y = \beta_0 + \beta_1 X$ can be built. Upon the production of the linear coefficients required to form the model, the model also produces standard error for the linear coefficients. For example, denote the standard error for β_1 to be SE(β_1). This process is internally understandable, because it allows end-users to establish hypothesis testing. An intuitive question raised is: is X important? To answer this question, an informative null hypothesis can be $\beta_1 = 0$, which implies that the independent variable X is not important. Alternatively, the coefficient $\beta_1 \neq 0$ which implies the model requires the existence of X_1. A t-test can be used and the t-statistics can be computed using $\frac{\beta_1}{\text{SE}(\beta_1)}$. If this test statistics is greater than certain threshold, i.e. for 95% confidence interval the threshold is at 1.96, then the null hypothesis can be rejected. There is no external effort required to understand a linear model. Hence, the linear model is considered interpretable.

The concept of explainability, however, requires more work. A deep neural network can be considered as a "black-box" model, of which its end-users are not required to understand the internal structure to be able to use it. The explainability of a deep neural network cannot be present unless some additional work is done. One famous algorithm is called Class Activation Map or CAM Zhou et al. (2016). In their work, the authors proposed an algorithm to utilize the internal layers from a deep neural network and extract the internal information to create the heat-map. This is an important step because these heat-maps can be overlaid on the original images to highlight the area of the image that the deep neural network used to make the final predictions. As such, the CAM technique are always done using the last convolutional layer of a deep neural network. A diagram of CAM is presented Figure 4.

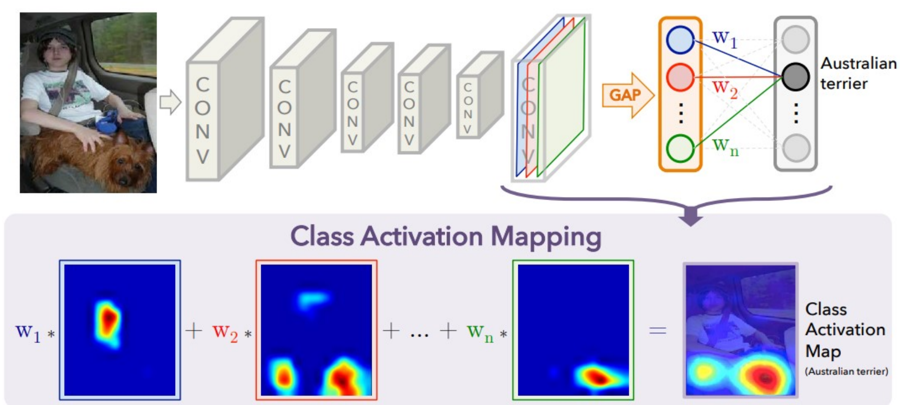

Figure 4: **Class Activation Map**.

A process such as using explainable method like CAM on a deep neural network to explain the decision making process can be considered explainable. In this case, the effort of developing CAM is external and is added on the existing deep neural network after the effect is taken. The deep neural network, despite its clear mathematical notation, can be very challenging to understand. With the external method, however, the deep neural network can be understandable to its end-users.

5 Future Work

From reading the author's work, it is apparent that the final scoring result is a function of a various of different features that is of certain moral values. It is quite challenging to develop a whole system to quantify these contributing factors. It is to my surprise that the system was actually developed at the first place. In the article, there was fruitful information about building up layers of definitions to support the ethics evaluation system. However, the discrepancy of these values and definitions via different culture is not mentioned at all. One potential upgrade for this ethical judgement model proposed by the author can be to implement certain variation according to different culture. However, this is difficult to implement, because this requires access of the data and features which are sometimes nearly impossible to collect and even more difficult to quantify.

Another aspect missing from the work is education. Education overall should go beyond high school or undergraduate degree though to quantify the data the degree might be a good way to start. To carry out detailed analysis, the first thing to discuss is whether education, from perspective of obtainable degree or not, should be added to the list of features at the first place. A good educated guess is that the sides supporting to add education to the system wins. If that is the case, the next question is to understand whether obtainable degree can be used to represent education. If not, then a good alternative should be raised in replacement of obtainable degrees. Moreover, when education is accepted as one from of ethics evaluation standard, the marginal association of how education affects the remaining features would then be the next challenge to tackle.

References

Badr, Y. and Sharma, R. (2022). Data transparency and fairness analysis of the nypd stop-and-frisk program. *ACM Journal of Data and Information Quality (JDIQ)*, 14(2):1–14.

Christman, J. and Zalta, E. N. (2015). The stanford encyclopedia of philosophy. *Autonomy in Moral and Political Philosophy [internet]. Edward N. Zalta ed.*

Cointe, N., Bonnet, G., and Boissier, O. (2016). Ethical judgment of agents' behaviors in multi-agent systems. In *AAMAS*, pages 1106–1114.

Lin, P., Abney, K., and Bekey, G. A. (2014). *Robot ethics: the ethical and social implications of robotics*. MIT press.

McLaren, B. M. (2006). Computational models of ethical reasoning: Challenges, initial steps, and future directions. *IEEE intelligent systems*, 21(4):29–37.

Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE intelligent systems*, 21(4):18–21.

Russell, S., Dewey, D., and Tegmark, M. (2015). Research priorities for robust and beneficial artificial intelligence. *Ai Magazine*, 36(4):105–114.

Yampolskiy, R. V. (2013). Artificial intelligence safety engineering: Why machine ethics is a wrong approach. In *Philosophy and theory of artificial intelligence*, pages 389–396. Springer.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929.