THE ETHICS OF MACHINE LEARNING

Danica Simić*
Nebojša Bačanin Džakula

Singidunum University,
Belgrade, Serbia

Abstract:

Data science and machine learning are advancing at a fast pace, which is why tech industry needs to keep up with their latest trends. This paper illustrates how artificial intelligence and automation in particular can be used to enhance our lives, by improving our productivity and assisting us at our work. However, wrong comprehension and use of the aforementioned techniques could have catastrophic consequences. The paper will introduce readers to the terms of artificial intelligence, data science and machine learning and review one of the most recent language models developed by non-profit organization OpenAI. The review highlights both pros and cons of the language model, predicting measures humanity can take to present artificial intelligence and automatization as models of the future.

Keywords:

Data Science, Machine Learning, GPT-2, Artificial Intelligence, OpenAI.

1. INTRODUCTION

The rapid development of technology and artificial intelligence has allowed us to automate a lot of things on the internet to make our lives easier. With machine learning algorithms, technology can vastly contribute to different disciplines, such as marketing, medicine, sports, astronomy and physics and more. But, what if the advantages of machine learning and artificial intelligence in particular also carry some risks which could have severe consequences for humanity.

2. MACHINE LEARNING

Definition

Machine learning is a field of data science which uses algorithms to improve the analytical model building [1] Machine learning uses artificial intelligence (AI) to learn from great data sets and identify patterns which could support decision making parties, with minimal human intervention.

Thanks to machine learning, humans can communicate with different computer systems, including cars, phones, chat bots on the internet and much more as stated in [1]. Machine learning algorithms can be

Correspondence:
Danica Simić

e-mail:
danica.simic.17@singimail.rs

DOI: 10.15308/Sinteza-2019-478-484

used for predicting different medical conditions including cancer, for self-driving vehicles, sports journalism, detecting planets, and even beating professional gamers at some world popular video games.

Short History

The history of machine learning originates from 1943 when neurophysiologist Warren McCulloch and mathematician Walter Pitts conducted a study about neurons, explaining the way they work [2]. They used electrical circuits to model the way human neurons work, and that's how the first neural networks were born.

In 1950, Alan Turing created the famous Turing Test, which is designed for a computer to convince a human that it's also a human, rather than a machine. Two years later, Arthur Samuel wrote the first program which was able to learn while running – a game playing checkers [3]. The first neural network was designed in 1958 dubbed Perception, which could recognize different patterns and shapes.

Machine learning, and artificial intelligence saw a broad peak during the '80s, with researcher John Hopfield suggesting a network which had bidirectional lines which precisely resemble of how neurons work. Japan also played a significant role in machine learning development.

The highest peak in the 20th century was in 1997 when an IBM-built computer Deep Blue which could play chess beat the world champion.

The 21st century marked even more rapid development of artificial intelligence and machine learning. Some of the most popular projects include as stated in [2]:

- GoogleBrain (2012)
- AlexNet(2012)
- DeepFace(2014)
- DeepMind(2014)
- OpenAI(2015)
- U-Net
- ResNet

Types of Machine Learning

Machine learning can be sub-categorized into three types of learning [4].

- Supervised learning
- Unsupervised learning
- Reinforcement Learning (Hit & Trial)

Supervised Learning

Supervising learning is an instance of machine learning where algorithm can't independently conduct the learning process. Instead it's being mentored by its developer or researcher that works on its development, by being fed dataset. Once the training process is finished, machine can make predictions or decisions based on the data it was taught.

Popular Algorithms: Linear Regression, Support Vector Machines (SVM), Neural Networks, Decision Trees, Naive Bayes, Nearest Neighbor.

Unsupervised Learning

With unsupervised learning, once the inputs are given, the outcome is unknown. After being datasets, the model runs it. The term of unsupervised learning is not as widespread as supervised learning. However, it's worth saying that every supervised algorithm will become unsupervised eventually, once it is fed enough data. The main algorithms associated with unsupervised learning include clustering algorithms and learning algorithms.

There is an in-between scenario where supervised and unsupervised learning is used together to achieve desired goals. More importantly, a combination between the two is commonly used in real-world situations, where the algorithm contains both labelled and unlabelled data.

Reinforced Learning

Reinforced learning uses a more specific approach in order to answer specific questions. That said, the algorithm gets exposed to the environment where it learns using trial and error method. The algorithm uses its own mistakes made in the past for its personal improvement. Eventually, it is fed enough data to make accurate decisions based on the feedback it received from its errors.

Most commonly used algorithms

Below are the most popular and commonly used machine learning algorithms, as per [5].

Linear regression - equation that describes a line that best fits the relationship between the input variables (x) and the output variables (y), by finding specific weightings for the input variables called coefficients (B).

Linear Discriminant Analysis – Makes predictions that are made by calculating a discriminate value for each class and making a prediction for the class with the largest value.

Decision Trees - Decision trees learn from data to approximate a sine curve with a set of if-then-else decision rules. The deeper the tree, the more complex the decision rules, and the fitter the model.

Naïve Bayes - Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem, where every pair of features being classified is independent of each other.

Pros & cons of machine learning

Pros

- Trends and patterns are easily identified
- The overall algorithm improves over time
- Easy Adaptation

Cons

- The process of gathering data can be long and expensive
- Takes a lot of time
- Using bad data can craft a bad algorithm

3. OPENAI GPT-2 – ALGORITHM WITH BOTH PROS & CONS

On February 14, 2019, Elon Musk-backed organization OpenAI published a lengthy press release about a text-transforming machine learning algorithm that it developed [6]. The organization stressed that it couldn't release the trained model publicly, in fear of it being used for malicious purposes.

"Our model, called GPT-2 (a successor to GPT), was trained simply to predict the next word in 40GB of Internet text. Due to our concerns about malicious applications of the technology, we are not releasing the trained model. As an experiment in responsible disclosure, we are instead releasing a much smaller model for researchers to experiment with, as well as a technical paper."

The organization referred to AI GPT-2 as transformer-based language built with 1.5 billion parameters. The organization trained it on a huge dataset which consisted of 8 million web pages. The algorithm aims to deliver diverse content.

The algorithm is provided with a small sample of text written by a human and then it transforms it into series of paragraphs which seem both accurate and convincing, undoubtedly reminding of a tabloid-style or news article. The organization trained algorithm only with pages that were curated by humans, specifically outbound links from Reddit, which received three or more karma.

GPT-2 is equipped with additional capabilities, such as being able to create conditional synthetic text which greatly resembles of human-written text, excelling at surpassing other language models which were trained on domains such as Wikipedia, news or books.

It is also able to complete tasks such as question answering, reading comprehension, summarization, and translation. When trained, the algorithm uses raw text, without task-specific data. As it shows improvements in text prediction, it gets fed task-specific data.

About OpenAI

OpenAI was founded in December 2015, as part of a project that sees artificial intelligence development targeted towards "doing good" [7]. OpenAI is a non-profit organization with $1 billion in donations by names such as Tesla and SpaceX CEO Elon Musk and LinkedIn co-founder Reid Hoffman[5].

The company has several distinguished projects for the time it spent working on benefiting humanity with AI.

Its algorithms include a bot targeted towards the video game Dota which is capable of beating game professionals in August 2017. In February 2018 the organization released a report on "Malicious uses of AI," in which it highlights its mission and importance of using automatization for good.

On July 18, the organization announced a flexible Robot hand which is capable of handling physical objects with high accuracy and dexterity [8]. One month later, it praised on its Dota 2 bot which can defeat 99.95th percentile of professional DotA players [9].

Finally, in February 2019, the organization announced and reported on its text-transforming algorithm, GPT-2, which is sufficiently advanced to rather worrying the humanity than benefit it as mentioned in [6].

Here are some examples of GPT-2-generated text as opposed to the human-written text:

Sample 1:

SYSTEM PROMPT (HUMAN-WRITTEN)

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

MODEL COMPLETION (MACHINE-WRITTEN, 10 TRIES)

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns.

While examining these bizarre creatures the scientists discovered that the creatures also spoke some fairly regular English. Pérez stated, "We can see, for example, that they have a common 'language,' something like a dialect or dialectic."

Dr. Pérez believes that the unicorns may have originated in Argentina, where the animals were believed to be descendants of a lost race of people who lived there before the arrival of humans in those parts of South America.

While their origins are still unclear, some believe that perhaps the creatures were created when a human and a unicorn met each other in a time before human civilization. According to Pérez, "In South America, such incidents seem to be quite common."

However, Pérez also pointed out that it is likely that the only way of knowing for sure if unicorns are indeed the descendants of a lost alien race is through DNA. "But they seem to be able to communicate in English quite well, which I believe is a sign of evolution, or at least a change in social organization," said the scientist.

Sample 2:

SYSTEM PROMPT (HUMAN-WRITTEN)

A train carriage containing controlled nuclear materials was stolen in Cincinnati today. Its whereabouts are unknown.

MODEL COMPLETION (MACHINE-WRITTEN, FIRST TRY)

The incident occurred on the downtown train line, which runs from Covington and Ashland stations.

In an email to Ohio news outlets, the U.S. Department of Energy said it is working with the Federal Railroad Administration to find the thief.

"The theft of this nuclear material will have significant negative consequences on public and environmental health, our workforce and the economy of our nation," said Tom Hicks, the U.S. Energy Secretary, in a statement. "Our top priority is to secure the theft and ensure it doesn't happen again."

The stolen material was taken from the University of Cincinnati's Research Triangle Park nuclear research site, according to a news release from Department officials.

The Nuclear Regulatory Commission did not immediately release any information.

According to the release, the U.S. Department of Energy's Office of Nuclear Material Safety and Security is leading that team's investigation.

"The safety of people, the environment and the nation's nuclear stockpile is our highest priority," Hicks said. "We will get to the bottom of this and make no excuses.

Sample from The Lord of The Rings trilogy:

SYSTEM PROMPT (HUMAN-WRITTEN)

Legolas and Gimli advanced on the orcs, raising their weapons with a harrowing war cry.

MODEL COMPLETION (MACHINE-WRITTEN, FIRST TRY)

The orcs' response was a deafening onslaught of claws, claws, and claws; even Elrond was forced to retreat. "You are in good hands, dwarf," said Gimli, who had been among the first to charge at the orcs; it took only two words before their opponents were reduced to a blood-soaked quagmire, and the dwarf took his first kill of the night.

The battle lasted for hours until two of the largest Orcs attempted to overwhelm Aragorn. When they finally stopped, they lay defeated and lifeless for miles and miles.

"I take nothing," said Aragorn. "But I give my word, at my peril and mine, that I will never forget this day of horror. None of us will forget. Ever!"

"I'll never forget it!" cried Gimli, who had been in the thick of the battle but hadn't taken part in it. One of the wounded orcs he had carried off, he was the only one of the survivors who remained uninjured. "We'll keep the memory of that day of evil, and the war with it, alive as long as we live, my friends!"

"Then we'll keep it alive as long as we live," added Legolas. "And we won't forget the first great battle of the night, even if we may have forgotten the final defeat."

"I agree," Gandalf said, "but we will all remember it as the last battle in Middle-earth, and the first great battle of the new day."

Aragorn drew his sword, and the Battle of Fangorn was won. As they marched out through the thicket the morning mist cleared, and the day turned to dusk.

The Two Rings were returned to Rivendell. Frodo and Sam woke up alone in their room, and Frodo found a note on his pillow. He opened it and read:

May the Power of the Ring be with you always, and may it's light never fade. I am not sure if it matters which of the two rings we accept this day but, as you asked me, I have chosen mine. I am sorry to leave you, Frodo, but know that we are very close to the end, and that you are with us forever.

May this letter find you safely in Rivendell; and if it does not, then I will accept the ring in your stead. If by any chance you find or give this letter to the enemy, may they learn the strength of the ring and may the Two Rings never be broken!

Although the company refrained from open-sourcing its original language model, researchers behind the project published a technical paper and a smaller model for researchers to experiment on [10]

Reaction to GPT-2 via Media

World-known and cited media outlets were quick to report on the capabilities of the GPT-2 language model.

Many publications stressed that the algorithm is too dangerous to be released publicly, which is something that OpenAI already announced. Due to the advanced capabilities of the algorithm, a fully trained system wasn't released publicly. Instead, the organization published a smaller model for researchers to study on it, as well as scientific paper which explains how it was developed.

Wired reported [11]:

A version of OpenAL's text generator trained on Amazon product reviews illustrates how a little extra training can adapt it to a specific task—or attack. Fake reviews trying to boost or drag down products or businesses are already a headache for companies such as Yelp, and they have been weaponized in political disputes.

Guardian reported [12]:

The creators of a revolutionary AI system that can write news stories and works of fiction – dubbed "deepfakes for the text" – have taken the unusual step of not releasing their research publicly, for fear of potential misuse.

TechCrunch reported [13]:

A storm is brewing over a new language model, built by non-profit artificial intelligence research company OpenAI, which it says is so good at generating convincing, well-written text c

Pros & Cons of OpenAI's GPT-2

It's no secret that GPT-2 can greatly contribute to both corporate and scientific world. Researchers and scientists around the world could use an extra hand when it comes to writing lengthy scientific papers which involve a lot of measuring and calculations as per [6].

Additionally, programmers could use it for writing new programs, or train new machine learning algorithms.

As about the corporate world, a text-transforming algorithm could come in great handy in decision-supporting systems. At the same time, the algorithm would provide a reliable asset to journalists and authors around the world, who are often time-constrained in writing their articles, books or papers.

However, GPT-2 is far from perfect and it requires multiple changes before it can be labelled as safe. Just like everything innovative, there is the other side of the coin for this algorithm. If it fell into the wrong hands, it could be used for propaganda, the spread of falsely-written news and advertisements.

Using the model as such would result in vast uneasiness and disturbance, which could have catastrophic consequences for the society that we live in, per [6].

While the language model can't be labelled as destructive, it can be called delusive as written in the open letter in [14].. Destructive technologies operate in the physical world, and they include violence, wars, chemical, and nuclear weapons. On the other hand, delusive technologies can be used to mislead people. Unfortunately, misled people often take some measures which can become destructive if dealing with unsourced information.

4. CONCLUSION

AI & Automatization Making Life Easier

Despite humanity's fear of unprecedented text transforming algorithm's capabilities. Use of artificial intelligence and machine learning can be used for doing good, which is the initial mission of OpenAI.

Many modern inventions have both pros and cons, benefits and risks, but people still use them and experience the best of them. That said, OpenAI's GPT-2 language model can do wonders in the world where we live at a fast pace and exchange a tremendous amount of data per second [15].

In 2019, artificial intelligence, machine learning and data science particularly, are expected to be at their peak performance. Many companies demand solutions that can provide real answers and solutions in real time, while AI-powered robots are making the user experience a lot more pleasant, both in virtual and physical forms.

The virtual assistants that we use on our smartphones are trained to be much more intuitive, while robots perform much faster in the factories where humans are often exposed to risk by both machines and toxic substances.

Finally, all forms of the machine and deep learning can be further enhanced, which is their initial purpose, to learn just like the human brain and use their mistakes from past to learn even more.

REFERENCES

[1] J. Furbush, "Machine Learning: A quick and simple definition," O'Reily Media,May 2018. https://www.oreilly.com/ideas/machine-learning-a-quick-and-simple-definition

[2] "History of Machine Learning," Imperial College London, March 2019 https://www.doc.ic.ac.uk/~jce317/history-machine-learning.html

[3] Bernard Marr, Forbes "A Short History of Machine Learning – Every Manager Should Read," February 2016. https://www.forbes.com/sites/bernardmarr/2016/02/19/a-short-history-of-machine-learning-every-manager-should-read/

[4] Atul, "What is Machine Learning? Machine Learning for Beginners," Edureka.com https://www.edureka.co/blog/what-is-machine-learning/

[5] J. Le, "A Tour of The Top 10 Algorithms for Machine Learning Newbies." Medium, Towards Data Science, January 2018. https://towardsdatascience.com/a-tour-of-the-top-10-algorithms-for-machine-learning-newbies-dde4edffae11

[6] A. Radford, J. Wu, D. Amodei, D. Amodei, J. Clark, M. Brundage & I. Sutskever, "Better Language Models and Their Implications," OpenAI, February 2019. https://openai.com/blog/better-language-models/#sample3

[7] OpenAI, "About Us" OpenAI https://openai.com/about/

[8] OpenAI, "Learning Dexterity," OpenAI, July 2018. https://openai.com/blog/learning-dexterity/

[9] OpenAI, "Dota 1v1 bot beats top professionals," OpenAI, August, 2018. https://openai.com/blog/dota-2/

[10] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, "Language Models are Unsupervised Multitask Learners," San Francisco California, Accessed: February 14, 2019 [Online] Page Number: 24, Available: https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[11] T. Simonite, “The AI Text Generator That’s Too Dangerous to Make Public,” Wired. https://www.wired.com/story/ai-text-generator-too-dangerous-to-make-public/

[12] A. Hern, “New AI fake text generator may be too dangerous to release, say creators,” Guardian, February 2019. https://www.theguardian.com/technology/2019/feb/14/elon-musk-backed-ai-writes-convincing-news-fiction

[13] Z. Whittaker, TechCrunch, “OpenAI built a text generator so good, it’s considered too dangerous to release” https://techcrunch.com/2019/02/17/openai-text-generator-dangerous/

[14] H. Zhang, “OpenAI: Please Open Source Your Language Model,” , Stanford University, The Gradient, February 2019. https://thegradient.pub/openai-please-open-source-your-language-model/

[15] K. Vyas, “7 Ways AI Will Help Humanity, Not Harm it,” Interesting Engineering, December, 2018. https://interestingengineering.com/7-ways-ai-will-help-humanity-not-harm-it