2022

BRIGHTWOLVES

The Pitfalls of Artificial Intelligence

Whitepaper

WWW.BRIGHTWOLVES.COM

THE PITFALLS OF AI

INTRODUCTION

Artificial Intelligence (AI) has rapidly emerged as a transformative technology, disrupting traditional industries and revolutionizing the way we live and work. However, as with any powerful technology, there are potential pitfalls and risks that must be considered. As AI continues to advance and become more ubiquitous, it is crucial to understand the potential risks and take steps to mitigate them.

In this series, we will explore 5 key pitfalls of AI. Every pitfall will be illustrated by a story and effective mitigation strategies will be discussed.

OVERVIEW OF ALL PITFALLS

1

Uninterpretable AI can lead to negative externalities

2

Micro-targeting using AI-generated subliminal messages is alarmingly efficient

3

AI models are approximations and can be tricked

4

Using unrepresentative training data leads to biased model

5

Recommendation algorithms are at the root of the polarization of our society

Pitfall 1: Uninterpretable AI can lead to negative externalities

Illustrative story

All big retailers use newsletters to promote their products. The data-oriented ones use AI models to build personalized newsletters with tailored discounts based on individual consumption data.

Target (a retailer in the USA), for example, started sending coupons for baby items to customers they predicted were likely to give birth soon. It worked so well that they got a complaint from a father accusing Target to encourage his underage daughter to get pregnant. However, a few days later, the father was surprised to learn that his daughter was actually pregnant.

While these AI-generated newsletters can boost sales, retailers may not anticipate potential negative externalities. For example, alcoholics might receive newsletters full of alcoholic products, anorexia patients might receive newsletters full of dieting pills and low-calorie products, and people with sugar addiction might receive newsletters full of sodas and ice cream.

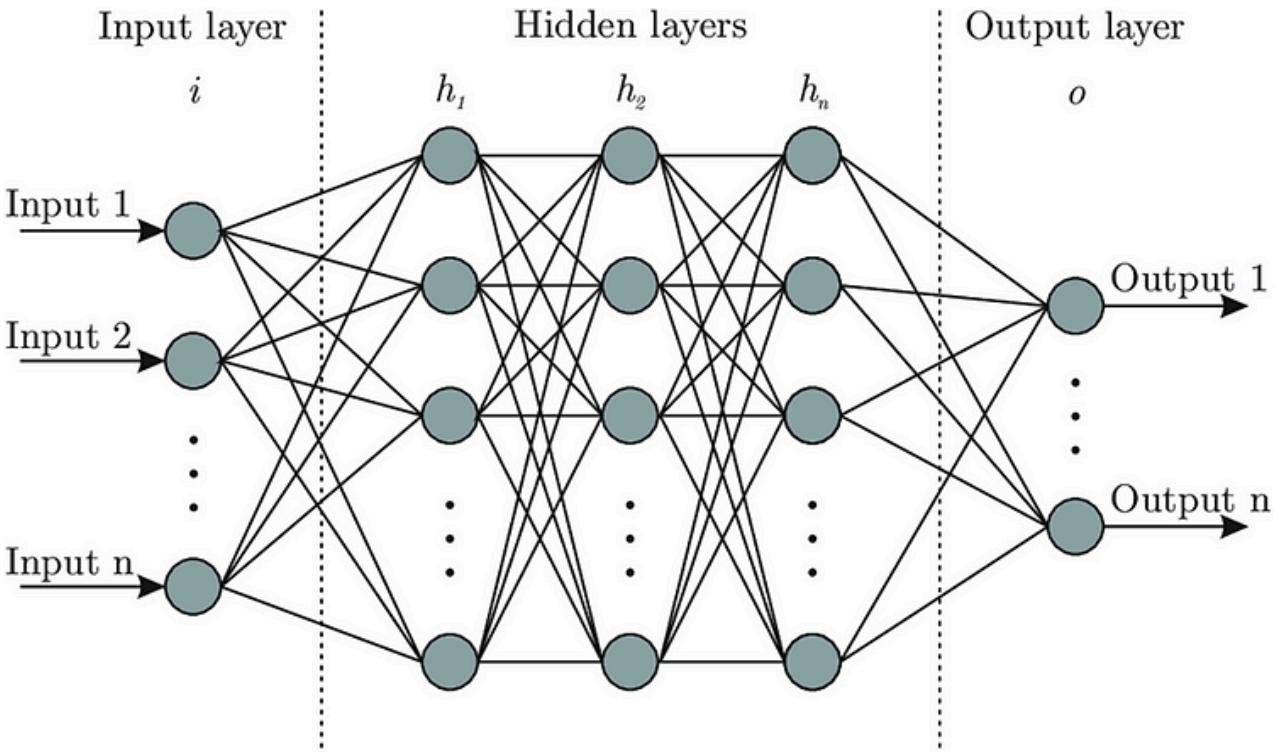
The AI models used for generating tailored content are often very complex. Which makes it very difficult for a retailer to determine why a particular product was suggested to a particular customer. More interpretable AI would make it easier for retailers to identify negative externalities.

Why?

In the last decades, machine learning models have become more complex due to the vast number of input parameters and the rising popularity of artificial neural networks (ANNs).

ANNs are efficient and data-hungry, but also opaque, earning them the "black-box" label. ANNs are based on biological neural networks, predicting output based on input data through a structure of connected nodes passing information with each other (see figure below). Due to continuous improvements in computing systems, ANNs are becoming increasingly complex, with some surpassing millions of parameters (such as Chat GPT-3, which is based on 175 billion parameters).

As the complexity of machine learning models will continue to increase, it is crucial to take into account the effects of the opaque nature of artificial neural networks.



How to mitigate?

Companies may find it challenging to work with black-box algorithms. However, the following strategies can be employed to minimize the risks associated with using uninterpretable AI.

1. **Develop explainability tools**
   SHAP, for example, lets you explain the output of any given machine-learning model. This increases the understanding, trust, and actionability of the trained models.

2. **Test, evaluate, and monitor algorithms**
   Thoroughly test the algorithm during its inception to ensure that it is making accurate and ethical decisions. Ensure some sort of human oversight when the model is deployed.

3. **Develop an internal AI ethics policy**
   Set up an organizational set of guidelines and principles to govern the development, deployment, and use of AI models. The policy should be governed by a cross-departmental team

4. **Upscale in-house data literacy**
   Generate awareness around the pitfalls of AI across the organization and train AI creators on the principles of AI (e.g., AI systems should be designed in a way that prioritizes simplicity over complexity whenever possible – Occam's razor)

5. **Consider interpretable algorithms**
   Many problems can be solved by a transparent and explainable AI model.

   Additionally, understanding how an AI model works will give you insights into your business process.

Pitfall 2: Micro-targeting using AI-generated subliminal messages is alarmingly efficient

Illustrative story

In 2016, Cambridge Analytica worked on a social media campaign for Donald Trump's presidential campaign. They used a highly successful micro-targeting campaign, using subliminal messages, and the rest is history. But how did they manage to achieve such a feat?

Cambridge Analytica gained access to the personal data of ~83 million Facebook profiles. Of these profiles, several hundred thousand users answered a series of personality questions that were used to create psychological profiles.
Using this data, a first model was created to infer the psychological profile of the remaining 82 million users based on their personal information. With these psychological profiles and personal data, a second model was created to select the most effective personalized ad for each user. The ads were deemed effective if they prompted Trump supporters to vote, discouraged Clinton supporters from voting, and swayed potential swing voters to the Trump side.

It is worth questioning whether all Trump voters voted for him due to the "raw" content of his program, or if the highly customized messages they saw on social media played a significant role. If the latter is more plausible, then AI could be a powerful tool for swaying public opinion. The person being swayed is often unaware he is presented with a specific framing of a message and why he has been selected for that message.

The ethical concern in this scenario revolves around the extent of human control in the presence of algorithms that possess the ability to comprehend an individual's emotional motivations to influence their thoughts, purchasing decisions, etc.

Why?

When taking a course on rhetoric and persuasion, one of the initial teachings is to understand your audience. Because depending on someone's socioeconomic background and psychology, an argument might resonate differently. Not necessarily due to its raw content but also due to the choice of words, its format, and so on.

For instance, if we were to persuade someone that flying is safe, our argumentation would vary greatly depending on the individual in front of us. For an analytical person, a statistical argument might suffice, such as: "Based on a scientific study, driving leads to 1.27 fatalities per 100 million miles driven compared to nearly zero per 100 million miles flown." Whereas for someone with a different psychological make-up, different arguments may be more effective. For example, "Uncle Marc is a pilot, and he is sure flying is safe" or "There is a seatbelt on the plane, so we should be safe."

If we know which arguments work for which personality types, we could very easily send out a tailored message to everyone to convince them that flying is safe. Identifying which message works for which personality type is something AI is very good at.

AI offers the potential to deliver personalized messages at scale. However, it is crucial to consider the ethical implications of using such technology, especially when it comes to persuasion and manipulation.

How to mitigate?

Micro-targeting using AI-generated subliminal messages can be a concerning issue for companies. Here are some strategies that can be used to mitigate this risk:

1. **Put the customer in the driving seat**
   Let your customer decide what degree of personalization in your ads he/she desires and what personal information he/she wants to share. This will foster trust and customer loyalty.

2. **Be transparent**
   Be transparent about how data is collected, stored, and used. This will reduce the likelihood of negative backlash.

3. **Monitor ad content**
   Let a person monitor the content of ads to ensure that they are not misleading or deceptive.

4. **Develop an internal AI ethics policy**
   Develop ethical guidelines for the use of micro-targeting and AI-generated subliminal messages who are compliant with data protection legislation (including GDPR).

By implementing these strategies, companies can mitigate the risks associated with micro-targeting using AI-generated subliminal messages. It is important for companies to prioritize user privacy and ethical considerations when using these technologies.

Pitfall 3: AI models are approximations and can be tricked

Illustrative story

To gauge the quality of education provided by universities, ranking models have been devised. These models output quality scores using inputs such as the number of papers published or the average salary of graduates of a university.

The number of papers published is a good indicator of the level of research but can easily be optimized by splitting one paper into five, publishing unfinished papers, or paying researchers to publish their papers under your university's name. Likewise, the average salary of graduates can be optimized by only accepting rich students or by reducing the number of students in the philosophy and psychology faculties and increasing the number of students in the engineering and business faculties.

Universities today are more and more dependent on these rankings and will put great energy and effort to optimize the model. These optimizations do not always improve the educational quality of the university and put pressure on other universities to follow suit.

When applying an artificial intelligence model, the same can happen. We should be aware that once the model is in place, people will understand which drivers influence the model's output and may seek to leverage this knowledge to their advantage.

Why?

Models approximate reality. IQ tests approximate intelligence, credit scores approximate individual financial solvency, and the GINI coefficient approximates inequality.

These approximations will never represent reality perfectly but are used to a great extent in society. When the inner workings of a model are public knowledge, it is easy to manipulate the model to predict a certain outcome.

All models are approximations. Essentially, all models are wrong, but some are useful. However, the approximate nature of the model must always be borne in mind.

George E.P. Box

How to mitigate?

AI models being approximations and being susceptible to being tricked can be a concern for companies. Here are some strategies that can be used to mitigate this risk:

1. **Regularly update models**
   Changes in a model's environment can quickly render a model outdated. Therefore, it is important to regularly update models with the latest available training data.

2. **Implement security measures**
   Implement security measures to ensure AI models are robust and protected from attacks. Adversarial testing, a technique to simulate attacks from malicious actors, can be used to identify vulnerabilities in AI models.

3. **Upscale in-house data literacy**
   Generate awareness around how to interpret and nuance AI models. Fostering a data-driven culture within your company will encourage collaboration and innovation around data analysis, ultimately leading to better company performance.

Pitfall 4: Using unrepresentative data training data leads to biased models

Illustrative story

Facial recognition software is known to work better for certain ethnicities and genders than others. The principal reason is not that specific ethnicities and genders are easier to identify but that certain ethnicities and genders were more heavily represented in the training data used to develop these algorithms.

For example, when Apple released its facial recognition system, Face ID, in November 2017, it was criticized for misidentifying Asian individuals. Similarly, a 2019 study found that SenseTime, a Chinese AI startup valued at over 10 billion dollars, misidentified Somali men 10% of the time. These examples highlight the need for diverse and representative training data when developing facial recognition software.

Why?

Machine learning is the sub-field of AI that encompasses all algorithms that learn from data. Once a model is trained, it can be used to predict an output based on previously unseen data. This process is known as generalization.

To ensure the model generalizes well or in other words performs similarly on the training data and unobserved data, the training data needs to be sufficiently representative.

Firstly, we need to ensure that the inputs and outputs of the training data accurately represent the phenomenon being modelled. For instance, consider a resume-scanner algorithm that assesses resumes and decides whether to proceed with a candidate or not. The training data inputs are all previous resumes received by a company and the output is the HR decision to proceed with the candidate. Let's now imagine the HR department was filled with racist, masochist, and islamophobia people who would always refuse a cv of non-whites, women, or people with an Arabic name. Our algorithm will model this behaviour and select candidates under the same criteria. Often models, despite their reputation for impartiality, reflect goals and ideology. When automating a process using past process data, we should always ask ourselves if the process was done well in the past. Secondly, we need to make sure that our training data is sufficiently complete. Does it include all types of unseen data?

This ensures that the model generalizes well on unseen data and does not become biased toward certain groups or attributes.

How to mitigate?

Using unrepresentative training data can lead to biased AI models, which can have serious consequences for companies. Here are some strategies that can be used to mitigate this risk:

1. **Use diverse training data**
   Use training data that is diverse and representative of the population being served.

2. **Check data quality**
   Ensure your training data is of good quality. Data of bad quality can only be used to build bad models.

3. **Upscale in-house data literacy**
   Ensure AI creators understand the different types of biases that can occur in training data. E.g., selection bias when certain data points are over- or under-represented or confirmation bias when the data is interpreted to confirm pre-existing beliefs.

4. **Understand the importance of a training, validation, and test set**
   Optimizing an AI model's performance on a validation and test set ensures it doesn't become too complex by fitting too closely to the training data, leading to poor performance when applied to new data.

5. **Conduct independent diagnostics**
   Conduct independent diagnostics to evaluate a model's accuracy and potential for bias. For example, by testing the model with dummy data not included in the training data.

Pitfall 5: Recommendation algorithms are at the root of the polarization of our society

Illustrative story

In the past years, TikTok has taken the world by storm. Primarily due to its addictive short-form video format and its algorithm. The algorithm analyses user behaviour, interactions, and preferences to tailor the content feed to each individual's interest. This personalized approach ensures that users are consistently served with videos that align with their tastes, leading them deeper into a highly curated world of specific content.

For the purpose of this paper, we conducted an experiment by creating a new TikTok account and deliberately engaging with 25 pro-life abortion videos. As a result, we observed a significant trend in the content recommendations. Out of the subsequent 25 videos that were suggested to us, 17 had a pro-life agenda and only 3 had a pro-choice agenda.

This experiment indicates that the algorithm has no incentive to propose viewpoints that are in opposition to the user preference.

Why?

In today's digital age, content providers, especially those in social media, are motivated to keep users engaged on their platforms for longer periods to generate more ad revenue. To achieve this goal, recommendation algorithms are often utilized. These algorithms suggest content that a user is most likely to engage with, be it an article to read, a video to watch, or a podcast to listen to.

However, this approach has its drawbacks. Users are often presented with content that aligns with their existing interests, resulting in them being trapped in their own comfort zone. For instance, a manga fan will be bombarded with manga-related content, and a soccer fan will be fed soccer-related content. While this seems harmless, things can take a negative turn for someone who is a conspiracy theorist or a far-right political supporter. They will be bombarded with content that confirms their beliefs.

On social media platforms, individuals can post content without being subjected to an objectivity check. However, if someone relies solely on social media as their primary source of information, they may fall prey to confirmation bias. As they may only consume content that confirms their pre-existing beliefs. Such behaviour is leading to an increasingly polarized society, particularly since the audiences of state-owned national "objective" news programs have been on the decline.

If we lived in a world where customized messages were ubiquitous in all aspects of our society, then each individual would potentially live in their own subjective version of reality.

How to mitigate?

Recommendation algorithms play a significant role in shaping our online experiences, but they can also contribute to the polarization of our society. Here are a few strategies that can be used to avoid polarization:

1. **Diversify recommendations**
   Recommendation algorithms should aim to provide users with a diverse range of content, including viewpoints that challenge their existing beliefs. Additionally, this can increase viewership by identifying new content categories that interest the user.

2. **Upscale societal data literacy**
   Our public education system should teach the wide public to think critically about the content they are presented with and to question the validity of the sources.

3. **Provide context**
   Recommendation algorithms should give objective context about the source, objectivity, author, political agenda, etc. of a recommendation.

4. **Use human oversight**
   Human oversight can ensure that recommendations are fair accurate, and diverse

By diversifying recommendations, encouraging critical thinking, providing context, and using human oversight, recommendation algorithms can be used in a responsible and effective manner that avoids the polarization of society.

Want to know more?

AI can bring tremendous value to an organization if it is well-managed and understood. However, implementing AI can be complex and time-consuming, requiring specialized knowledge and resources.

At BrightWolves, we specialize in providing customized advice and solutions tailored to specific business needs. Our expertise in AI can help accelerate your digital & data transformation by providing valuable guidance on best practices and implementation strategies.

What sets us apart is our focus on the business side of data analytics, rather than just the technical aspects. We understand that data is only valuable if it helps businesses make better decisions and achieve their goals.

If you want to know more, do not hesitate to reach out to our AI experts:

Olivier De Moor

BRIGHTWOLVES