The Future of AI; Bias Amplification & Algorithmic Determinism

Posted on 17th July 2019

Diversity | Emerging Tech

Written by Jillur Quddus, Data Scientist, Methods

Technology provides us with the toolkit to change lives for the better – but if unchecked, it also has the power to discriminate and reinforce stereotypes and bias.

This is true of any technology past, present and future. But it is the advancement of distributed systems capable of storing and processing massive amounts of previously disparate data coupled with the emergence of artificial intelligence (AI) into our everyday lives that requires us to urgently refocus on how technology is being architected, engineered, tested, deployed and governed, to ensure that its impact remains positive – and only positive.

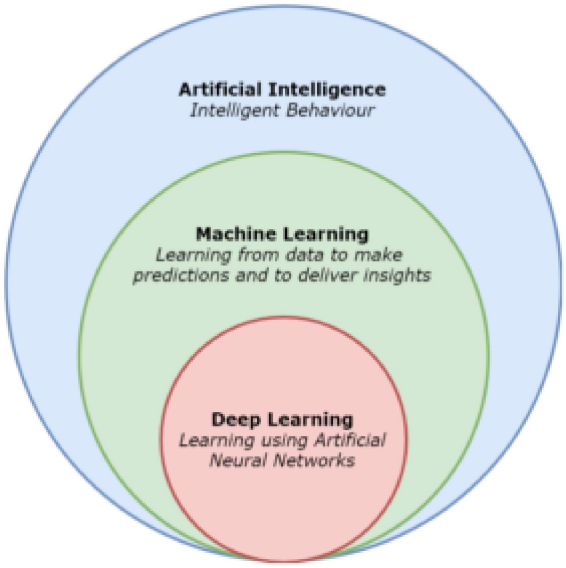
First Principles

Whether we explicitly asked for it or not, the fact is that, today and right at this very moment, artificial intelligence is impacting your life. The majority of the time, AI is invisible to us – whenever we enter a search term into Google, visit websites, use social media, browse Netflix, use online banking, use public transport or walk the streets of any major city, AI algorithms are busy working away in the background to return relevant search results, analyse your browsing habits, social network, viewing preferences, detect signs of fraud and to capture an image of, and process, your face. With such a deep and far-reaching, yet paradoxically invisible, global footprint, the dangers of AI if unchecked and not properly governed are significant and tangible, not least to the groups of people it will adversely affect (usually those already at a social disadvantage).

But before we being to explore these dangers, let us go back to first principles and definitions. AI is a broad term given to the theory and application of machines that exhibit intelligent behaviour. Machine learning (ML) is an applied field of study within the broader subject of artificial intelligence that focuses on learning from data by detecting patterns, trends, boundaries and relationships in order to make predictions and ultimately deliver actionable insights to help decision making. A fundamental tool used in machine learning is that of probabilistic reasoning – creating mathematical models that help us to draw population inferences to understand or test a hypothesis about how a system or environment behaves. As such, machine learning is data-dependent and, supervised and unsupervised machine learning both make an implicit assumption that trends and patterns identified in historic data in order to learn a mathematical function, can be used to make inferences and predictions about the future using that same mathematical function.

Deep learning

(DL) is a sub-field within machine learning where the goal remains to learn a mathematical function that maps inputs to outputs. However, deep learning learns this function by employing an architecture that mimics the neural architecture found in the human brain in order to learn from experience using a hierarchy of concepts or representations, and gradient-based optimisation algorithms. Deep learning is still data-dependent however, and the more data used to train a neural network the higher the predictive accuracy.



Bias Amplification

Now that we have an understanding of what artificial intelligence is, it is clear that data – and lots of it – is required for machine learning to be effective. The more data, the better the predictive powers of the resultant model.

But what if the data itself contains bias? In this case, the model itself will not just perpetuate that bias but in many cases will amplify it. So why do machine learning algorithms amplify bias? To answer this question, we must revisit the mathematical nature of their training structure.

Recall that the goal of machine learning algorithms is to learn a mathematical function to map input data points to outputs. Normally these outputs take the form of a value or classification. In order to learn this mathematical function and to maximise its accuracy, machine learning algorithms seek to minimise the number of errors it makes by reducing the variance in the predictions made by the final model. To reduce variance, algorithms introduce assumptions, referred to as bias in the model, to make the function easier to approximate. By introducing more bias, there will be less variance in your final model, but at the expense of greater divergence from the real-world reality in which your model

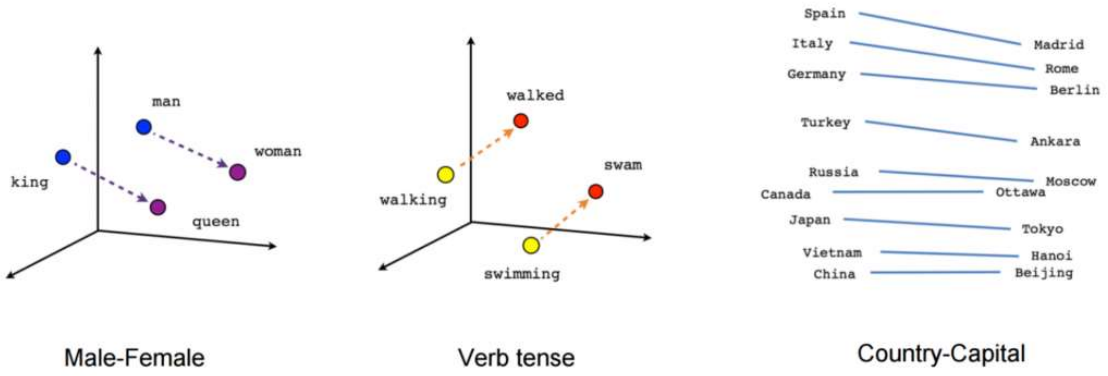
operates. But not enough bias, and the predictive accuracy of your final model is reduced as there is greater variance. As such, many machine learning algorithms expose a bias-variance trade-off through the manipulation of hyper-parameters so that the developer or data scientist may control this balance.

However therein lies a problem – if a goal of machine learning is to train models to maximise their predictive accuracy, then any bias in the data that feeds these algorithms will be amplified due to the algorithm generalising the data in order to reduce variance. Therefore if the data contains gender or racial bias for example, this bias will be preserved and amplified to help make the model more accurate, perpetuating a vicious circle of bias and discrimination.

Gender bias in Natural Language Processing

Let us take a look at an example algorithm to highlight how the problem of bias amplification is a physical reality in today's artificial intelligence-driven world. Natural language processing (NLP) refers to a family of computer science disciplines, including information engineering, linguistics, data management and machine learning, with the goal of helping computers to understand natural language used in speech and text. Natural language processing is employed in a variety of scenarios, from interaction with chatbots and virtual agents such as Alexa, to foreign language translation and understanding and classifying written text such as legal contracts and health reports.

One common family of feature-learning techniques used in natural language processing is called word embeddings. A word embedding represents each word or phrase found in a given collection of texts as a vector. Words with similar semantic meaning will have vectors that are close together, and the relationship between words can be quantified through the vector differences. As such, word embeddings are an extremely useful tool for a wide variety of prediction tasks related to natural language. For example, when we wish to determine sentence similarity, such as when asking a chatbot or Google a question, it will employ word embeddings to see if that question has been asked before and take action accordingly based on a similarity index. Another example is analogical reasoning where we wish to predict semantic relationships. A widely used implementation that learns vector representations of words is word2vec which provides semantic relationships between words as follows:



Male-Female Verb tense Country-Capital

A version of a word2vec model was trained by Google engineers on a dataset of hundreds of millions of Google News articles, containing hundreds of billions of words in total. What happens when we apply these trained word2vec word embeddings to analogical reasoning tasks? For example, take the following analogical phrase:

Dog is to **Puppy** as **Cat** is to _____?

The Google-trained word2vec embeddings predicts "**Kitten**". Let us look at another example:

Man is to **Computer Programmer** as **Woman** is to _____?

The trained word2vec embeddings in this case predicts "**Homemaker**". This is a clear example of how gender bias in the data used to train a model (in this case Google News) perpetuates through to the final classifications and predictions made. A demo web application that uses the word2vec model trained by Google on the Google News dataset may be found at https://rare-technologies.com/word2vec-tutorial/.

Word embeddings are a notable example of bias perpetuation and amplification, exacerbated by the fact that their usage is extremely common in tools developed by large technology companies and which we use every day. **Google Translate** is an example of this – a tremendously useful tool but which uses word embeddings and hence can preserve gender bias.

For example, enter the following English phrase into Google Translate and select a target language that has no gender-distinction pronouns such as Finnish or Hungarian:

English	*Finnish*
He is a nurse. She is an astronaut.	Hän on sairaanhoitaja. Hän on astronautti.
	Hungarian
	Ő egy nővér. Ő egy űrhajós.

If you then use Google Translate again to translate the Finnish or Hungarian back into English, you get:

She's a nurse. He's an astronaut.

You will see that the gender pronouns have been swapped around, helping to reinforce gender stereotypes.

Algorithmic Determinism

Recall that both supervised and unsupervised machine learning make an implicit assumption that trends and patterns identified in historic data in order to learn a mathematical function can be used to make inferences and predictions about the future using that same mathematical function. Personalised recommendations from services like Netflix and Amazon, and personal assistants and virtual agents like Alexa and Siri, all work using this same fundamental principle of **algorithmic determinism** – that what has happened before can be used as a context in which to predict and recommend actions in the future.

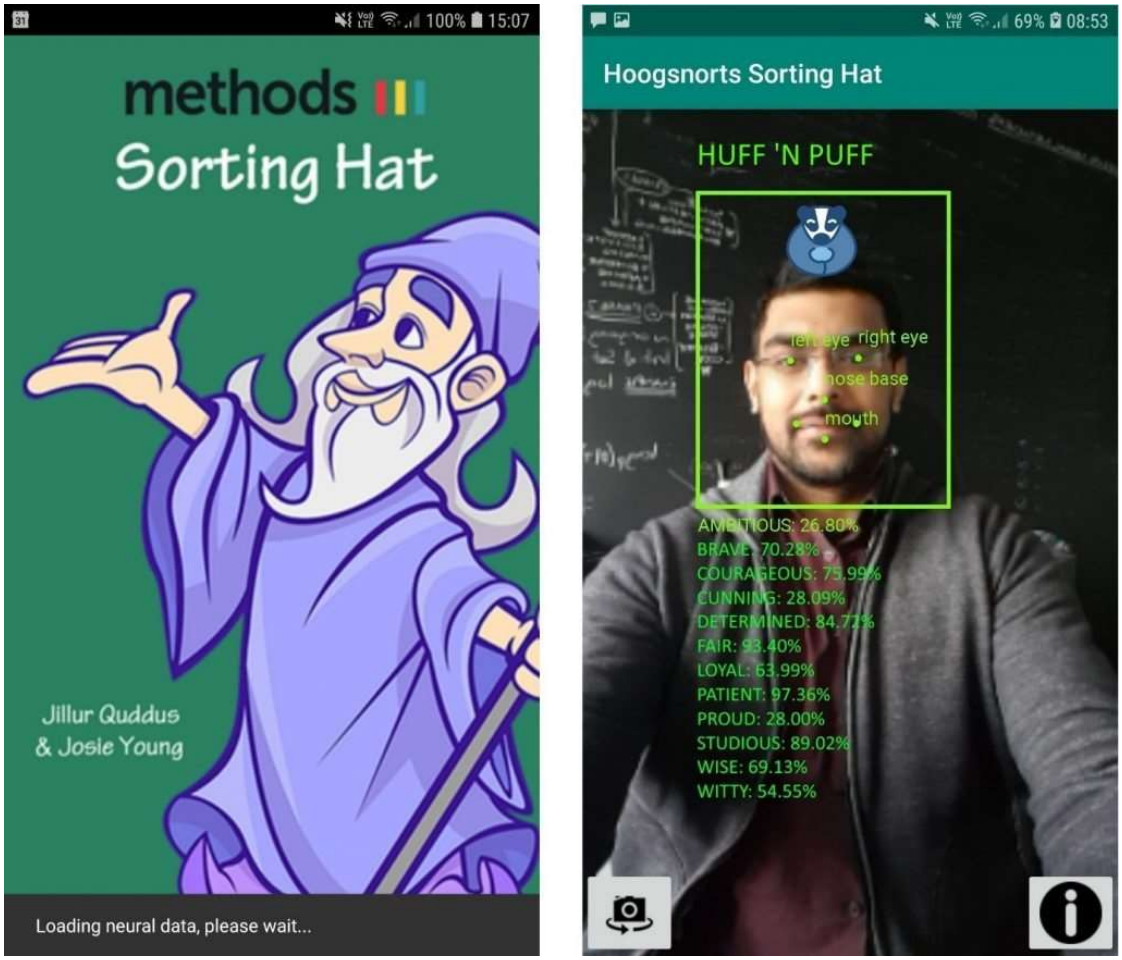
However, the real danger of the inferences and predictions provided by these machine learning models is that as they provide increasingly deterministic recommendations, they perpetuate bias, stereotypes and discrimination by reinforcing historic and existing beliefs. We will continue to watch only those shows on Netflix, or continue to read only those books on Amazon or tweets on Twitter, whose content we liked before, rarely or even never to expand our horizons, open our minds to new ways of thinking nor meet people and engage with content we may initially disagree with. Artificial intelligence then

becomes a tool with which to maintain existing behaviour and beliefs, inflexible to behavioural change. And this is inherently dangerous to any free society – where your past dictates your future.

Sorting Hat

To highlight the amplification of bias and algorithmic determinism in artificial intelligence in a light-hearted manner, Methods' Emerging Technology (ET) practice recently developed an Android app called the "**Hoogsnorts Sorting Hat**". Inspired by the sorting hat from Harry Potter, we developed an end-to-end deep learning neural processing pipeline coupled with an Android augmented reality app designed to showcase the applied end-to-end integration of machine learning, deep learning and augmented reality whilst highlighting the ethical and social impact of artificial intelligence as part of its ever-increasing prominence in decision making in modern society. The end-to-end neural pipeline can be split into two primary parts:

1. **Training the neural classifier** – the Chicago Face Database (CFD) was used to generate the training data in a crowd-sourcing style event. Staff from Methods' Emerging Technology (ET) practice manually classified every face as to whether they believed the person in question possessed subjective personality traits, resulting in binary (Yes or No) features for each subject. A convolutional neural network was then trained using the Deeplearning4j Java library, employing a stochastic gradient descent-based optimisation algorithm and a multinomial logistic regression/softmax-based activation function in the output layer.
2. **Neural Classification via the Android Augmented Reality App** – the Hoogsnorts Sorting Hat Android App uses Google's Vision API to automatically identify human subjects and their faces using Deep Learning. Upon identification, the Hoogsnorts Sorting Hat Android App then uses the device's camera to take a picture using the boundaries as identified by the Vision API. That picture is then sent to the trained convolutional neural network in real-time to calculate probabilities for each subjective feature. The probabilities are aggregated using a scoring engine which then makes the final determination as to which Hoogsnorts house the subject should belong to. The available houses are "Ovendoor", "Blytherin", "Pandapaw" and "Huff 'n Puff". Both the predicted feature probabilities and the overall school classification are then sent back to the Android app and overlaid on the subject using basic augmented reality.



Methods' Hoogsnorts Sorting Hat

Whilst tongue-in-cheek, the app does serve to highlight the dangers and inherent flaws of artificial intelligence if improperly designed and governed. The dangers of classifying people based on flawed and biased data are equally pertinent to every-day scenarios such as applying for insurance online and resultant premiums based on subjective features open to discrimination, using cameras to take photographs of people in an attempt to identify criminals and racial bias, digital recruitment and gender bias, and recommendation systems and political bias.

The Future

Artificial intelligence has the power to positively transform society in a way very few technologies can, and inspire future generations through its seemingly limitless potential – as demonstrated in my previous articles introducing Quantum Deep Learning and Probabilistic Programming Languages. The applications of artificial intelligence are bounded only by our imaginations, and the underlying mathematics driving the latest research and developments is both beautiful and mind-blowing. However as with all such things that have the power to affect people's lives, it can also be

exploited and misappropriated – in the case of artificial intelligence however, the dangers are just as profound even when the intention is good, through the unwitting perpetuation and amplification of bias and discrimination.

Sadly, bias and discrimination is a way of life for many people and groups of people, and has been for hundreds and probably thousands of years. So whilst eradicating something so deeply entrenched into society may take generations, we can take tangible steps *now*to ensure that any and all data-driven systems that we develop, and have developed, minimise bias perpetuation and amplification. This includes designing and enforcing, through effective policy and governance, fully-transparent and accountable algorithms coupled with adherence to ethical frameworks that seek to mandate governance and accountability at each stage of the data life-cycle. Finally, and arguably most importantly, we must strive to improve diversity in the fields of mathematics, science, engineering, technology and philosophy. By improving diversity in these fields along the entire academic and professional chain, we allow for all areas of society to have its say on the future relationship between humans and artificial intelligence.

Originally posted at Methods.

Image from UKBlackTech.

MORE THOUGHT LEADERSHIP ▷

Leave a Reply

Your email address will not be published. Required fields are marked *

Comment

Name *

Email *

Website

Save my name, email, and website in this browser for the next time I comment.

Post Comment

RELATED STORIES

Gender inequality in AI

July 2019

AI and Cognitive Computing

June 2019

The Trampery Old Street, 239 Old St, London EC1V 9EY

| press@digileaders.com