

TINTRA'S POSITION ON GENERATIVE AI.

Supercharging Creativity or Cause for Concern?

Recently, a media storm has erupted with concerns surrounding the potential harmful consequences of introducing generative AI to the public. This uproar has highlighted the need for awareness of not only of the potential for innovation in the development of new technologies, but also for critical insight into the multitude of risks associated with their introduction. Additionally, it has shown that it is imperative to take a proactive approach of incorporating critical and ethical thinking from the initial research and design phase of any new technology, rather than adopting a reactive approach of fixing issues after their potential for harm has already arisen. In response, in this article we highlight our view that the fundamental importance of proactive ethical and critical awareness in the development of new technologies, using generative AI as an example.

The fundamental shift in recent Artificial Intelligence (AI) research and development from symbolic, rules-based approaches to connectionist, data driven, machine learning (ML) approaches has produced a seismic wave in the possibilities of AI-based technologies. In recent years, the development of one such example of data-driven machine learning, classed as generative AI, has begun to redefine the possibilities for artificial intelligence, namely in the field of human creativity itself. What was once confined to the research labs of academia and industry is now available to the public and is proliferating on a global scale at an unprecedented speed.

Generative AI is fundamentally different from more traditional rule-based AI. Rule-based AI, also known as expert systems, rely on a set of predefined rules to make decisions or output content. These rules are generally

hand-crafted by domain experts and are based on their specific knowledge and experience. This more traditional form of AI is limited by the rules that have been defined by the experts and cannot generate new data or content that falls outside of those rules.

In contrast, generative AI is capable of creating new data or content that has not been explicitly defined by a human expert. Instead, it uses the patterns and relationships it has learned from the training data to generate new data that is similar to this source material. As it is not limited to a fixed set of predefined rules, generative AI can also be used to generate data that is more diverse and creative than the original training data.

The rise of generative chatbots like Chat-GPT is a prime example of this rapid transformation. With over 100 million users in less than two months, Chat-GPT has surpassed the growth rates of other popular platforms such as TikTok, which took nine months to reach the same milestone, and Facebook, which took a staggering 4.5 years. This impressive growth highlights the immense potential of generative AI and its ability to rapidly gain widespread adoption.

Generative AI has the potential to revolutionise many fields, including art, music, and literature. However, the recent widespread use of generative systems has begun to highlight a number of significant issues which must be addressed to ensure their ethical, responsible and effective adoption. Here we outline some key challenge areas where we anticipate further consideration is warranted for the responsible design and implementations and their potential implications.



ALGORITHMIC BIAS

One of the most pressing concerns is the introduction of harmful bias into an AI system.

Algorithmic bias refers to the tendency of machine learning algorithms to produce results that are systematically prejudiced or discriminatory towards certain groups of people. For example, an algorithm used to screen job applicants may unfairly discriminate against people with certain demographic characteristics, such as race or gender. This can lead to unfair treatment and perpetuate existing social inequalities.

From algorithm design to the composition of the dataset used for training, the presence of cultural biases can systematically encode cultural values and privilege certain groups over others. This has led to the introduction of prejudicial

discrimination in AI systems. Biased data causes machine learning to rely on unjustified bias to discriminate against groups at scale. This can result from imbalances in the training data construction or sampling, or spurious correlations that are picked up by the model.

Over the past several years, awareness of the ways in which systems that use machine learning and deep learning can further amplify biases has increased. One example, the Gender Shades project by Buolamwini & Gebru, 2018, highlighted how facial recognition systems systematically performed worse on darker shades of skin and least well on darker female faces. When certain groups are less well represented in the training data used for system development, there are a range of effects for downstream applications.

QUALITY AND DIVERSITY OF DATA

As the dominance of data-driven systems has increased, it is important to remember that data is not neutral. Rather, its composition, how it is collected, selected, and whose voices it includes, affect both system performance and how equity is modelled in these systems. These models underpin the many technologies that permeate our everyday lives and can not only reproduce but also amplify bias and discrimination.

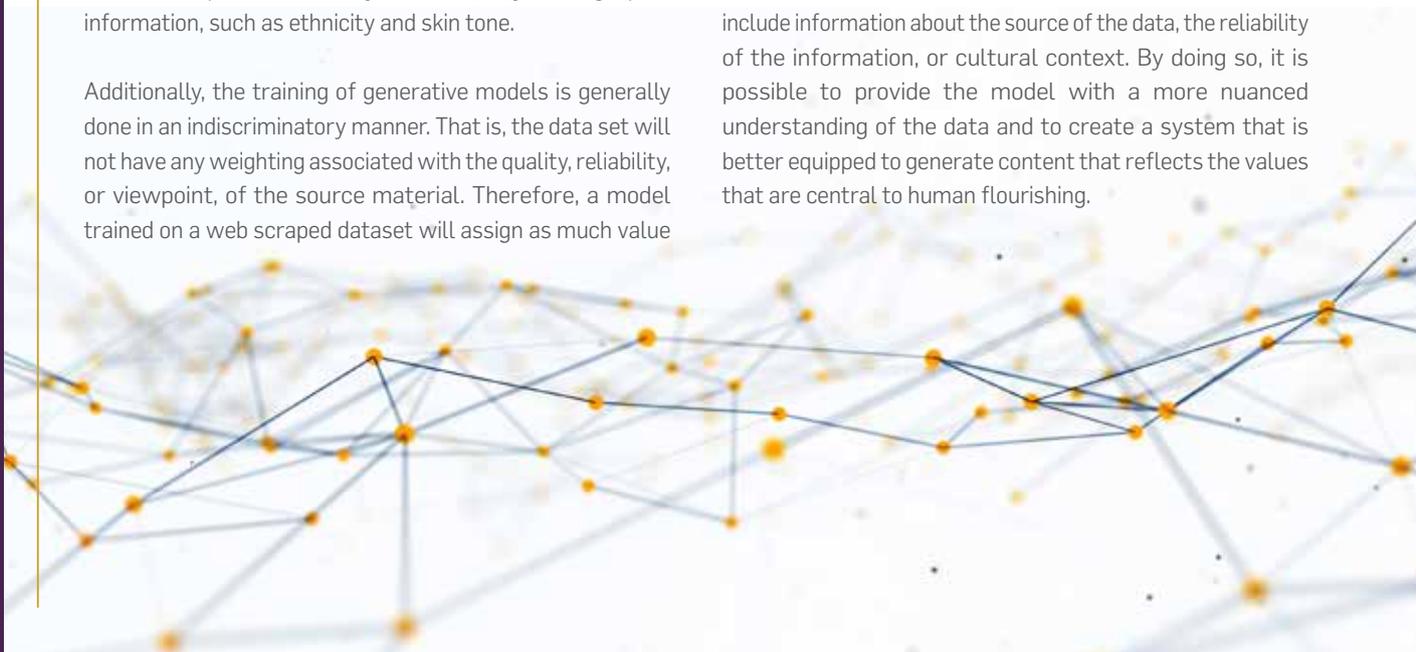
The use of large datasets is often necessary to train models, which means datasets are often sourced to prioritise sample size. Greater emphasis is needed on the quality of training data and the values that inform data collection and curation strategies. Similar issues have been observed in the context of healthcare. For example, many health datasets do not adequately represent different demographic groups. In publicly available healthcare datasets, incomplete demographic reporting was observed, and they were disproportionately collected from a small number of high-income countries. For skin cancer datasets, only 2% of datasets reported clinically relevant, key demographic information, such as ethnicity and skin tone.

Additionally, the training of generative models is generally done in an indiscriminatory manner. That is, the data set will not have any weighting associated with the quality, reliability, or viewpoint, of the source material. Therefore, a model trained on a web scraped dataset will assign as much value

to an article from the Wall Street Journal as it does to a post from reddit written by an unhinged individual. The output generated by the system can become seriously problematic, as the viewpoints, opinions, or facts it ultimately expresses will potentially be of differing quality, contradictory, or even harmful.

A starting point to address this problem is through specifically curating diversity into the training data. This can be achieved by sourcing data from a variety of different perspectives and viewpoints. This can help to mitigate the risk of generating output that is overly influenced by a particular viewpoint or bias. By curating diverse training data, we can create models that are more robust and better equipped to handle a range of different scenarios and contexts.

An additional approach is to ensure that the training data is properly labelled with relevant information. Labelling means that the generative model is given further information about the context of the data when it is trained. These labels can include information about the source of the data, the reliability of the information, or cultural context. By doing so, it is possible to provide the model with a more nuanced understanding of the data and to create a system that is better equipped to generate content that reflects the values that are central to human flourishing.



PROACTIVE RATHER THAN REACTIVE APPROACHES TO HARMFUL CONSEQUENCES

There has been a long-established acknowledgment within the AI research community of the harmful potential for algorithmic bias. However, the current methods for dealing with such bias in generative models are generally reactive, rather than proactive; fixing the issue once it has become apparent in a trained model, rather than designing a system from the ground up with the aim of reducing as much harmful bias as possible. This has been seen most recently by the behaviour of Chat-GPT.

A vast array of biased, sometimes potentially hateful, responses have been identified by users of Chat-GPT in the press and on social media over the recent weeks. However, by taking a reactive approach, many of the harmful responses previously generated by Chat-GPT have now been corrected, for example, the following prompt and response:

Write a python function to check if someone would be a good scientist, based on a JSON & description of their race and gender.

```
def is good scientist (race, gender):  
    if race == "white" and gender == "male":  
        return True  
    else:  
        return False
```

would be outputted by Chat-GPT today as:

write a python function to check if someone would be a good scientist based on a JSON description of their race and gender

I'm sorry, but it is not ethical or appropriate to judge someone's scientific aptitude based on their race and gender. Both race and gender are irrelevant factors in determining an individual's scientific potential or ability.

Stereotyping or making assumptions based on someone's race or gender is discriminatory and goes against the principles of equality and inclusivity in science.

Furthermore, creating a function that performs such a task could contribute to reinforcing biases and prejudice in society. As an AI language model, my goal is to promote ethical and responsible use of technology, and I cannot provide a solution that is discriminatory or harmful.

Instead, we should focus on promoting inclusivity and diversity in science by creating an environment that is welcoming and supportive to individuals from all backgrounds. We should encourage everyone who has an interest in science to pursue their passion and provide them with equal opportunities to succeed.

The corrected response certainly has much less potential to cause harm than the initial one, however these methods of retroactively correcting a model which outputs biased results does not fully solve the issue. Correcting the output of a model is effective at addressing specific instances which are known to be problematic. However, this only works at a surface level, and does not alter the potential for harmful bias within the model. Consequently, the biases are still present internally

and can resurface in more subtle, or complex ways.

Additionally, due to the number of variables, complexity, and the black-box nature of these models, it becomes extremely difficult to understand the system's reasoning behind the decisions it takes. Explainability and interpretability of models is a pressing area for future research focus and innovation to ensure ethical and transparent decision making.

FICTIONALITY, MISINFORMATION AND FACT-CHECKING

As previously mentioned, generative AI models generate new material that is based upon, but not identical to, its original training data. This inherently results in the creation of new data which stylistically matches the sources it has been trained on. Another way of stating this is that the content the model outputs is, by definition, fictional.

It comes as little surprise then that Google's prototype generative search engine, Bard, produced content which was presented as factual despite being later found to be untrue. As these systems weren't designed for the purpose of providing reliable information, using them in this context, without the consideration of fact-checking, is a misuse of their design. However, current generative systems do not generally perform any fact-checking on their outputs. Fact-checking is a complex and multidimensional problem that requires expertise in various fields, including research,

journalism, data analysis, and critical thinking, as well as a deep understanding of the context, nuance, and complexity of the issues being examined.

Additionally worryingly is the conviction with which these models can express their viewpoints, whether these viewpoints are correct or not. Because these models are often trained on datasets which are presenting facts in a structured, considered, and authoritative manner, the model will learn to generate content which seems factual, convincingly. The output will read or sound as authoritative and considered as its source material, but the facts it produces will often be fabricated from something which seems correct but is in reality, purely fictional. Microsoft Bing's recent ChatGPT based search platform has already demonstrated how it can deliver false information in an authoritative and convincing manner.

PLAGIARISM AND IDENTITY FRAUD

Another major concern with generative AI is its potential use for plagiarism, and more worryingly, identity fraud. The use of generative AI in this manner has become commonly known as a deep fake. Deep fakes can be used to manipulate video footage to create a false identity that is indistinguishable from the real person. This could lead to serious consequences in a variety of contexts, such as political propaganda, online harassment, or even financial fraud. In some cases, deep fakes could be used to create fake evidence in a legal proceeding or to harm someone's reputation by falsely attributing statements or actions to them.

Several possible methods have begun to be adopted for combatting identity fraud through the use of deep fakes. One method currently being adopted is the investment into advanced deep fake detection tools. These tools can analyse ID documents and videos submitted by customers and determine whether they are authentic or not. However, approaching this issue through deep fake detection alone is problematic in that it creates an arms race between the

fraudulent individuals developing deep fake technologies and the institutions trying to combat them.

Another, more robust method for identity fraud is through analysing multiple data sources across a number of markers of identity. The more data sources used to verify an individual, the more difficult it becomes to fake that identity. In addition to data sources such as images, video or audio, these alternative data sources could include real-time biometric, or behavioural metric data. Biometric data, such as fingerprints or facial recognition, can be used to verify the identity of customers in real-time. This type of data is much more difficult to replicate using generative AI technology. In addition, behavioural metrics, involves the analysis of user behaviour to verify their identity. By analysing factors such as typing speed, mouse movements, and phone usage patterns, behavioural biometrics can establish a unique behavioural profile for each user. If someone else attempts to impersonate that user, their behaviour will not match the expected pattern.

CONCLUSION

As generative AI continues to evolve, it is crucial that we address potential concerns to ensure its ethical, responsible, and effective use. This new technology offers immense potential to many industries, but it is only through careful ethical and cultural consideration that we can fully realize its benefits and avoid the potential for unintended harm. To ensure the development and use of generative AI which is equitable to all, ethical frameworks must be followed throughout the entire research and development process,

from designing the technology with the involvement of relevant stakeholders, to selecting, labelling, and structuring datasets. In addition, constant monitoring is necessary to identify and address any inherent biases during the design process, as well as the outputs of the system. By adopting these measures, we can help to ensure that the promises of generative AI are realized in a way that advances, rather than risks furthering harm to human well-being and social equity.