

# Ethical Implications of Generative AI

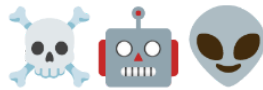










Table of Contents

<b>Ethical Implications of Generative AI</b>	<b>1</b>
Table of Contents	2
Introduction 	3
Agenda 	3
Literature Review 	3
Types of Bias/Risks 	3
Mitigation Mechanisms 	5
Key Takeaways 	5
Conclusion 	6
References 	6

## Introduction

The rise of Generative AI has led to a Cambrian explosion of intelligent conversational agents and support assistants. This proliferation of LLMs has opened a Pandora's box because of the challenges it creates for all stakeholders in the ecosystem - individual users, corporations and government organisations/regulators.

The purpose of this session is to gloss over the rise of LLMs and how easy access to Web APIs has spurred an explosion of applications using such models. Then we will delve into details of how we at eyeo ensure that the models that we develop at web-scale are not biased towards certain characteristics of the input data. Finally, we discuss what the ethics of adopting these technologies will look like - what are the challenges and who are the stakeholders because at the end of the day, it's not only an Engineering problem but also a social and economic problem.

## Agenda

- History and Rise of LLMs - How we got here?
- Where do we stand and how the web has spurred greater adoption of Generative AI
- Types of Biases in LLMs
- How we at eyeo take concrete steps to mitigate the bias in our models
- How do we define ethics to start with -
  - How do we make ethics more objective rather than subjective? How can it be a part of the engineering development process (i.e. incorporate a set of best practices)?
  - Is it a reflection of how the society is today or how we envision it to be?
  - What are the societal changes ushered by Generative AI that regulation can help mitigate?
  - Who should be the gatekeepers when it comes to defining Ethics for Generative AI and AI in general?

## Literature Review

The history of LLMs is rooted in the Transformer architecture. The architecture was introduced by researchers from Google in their seminal paper titled - "Attention is All You Need". It introduced multi-head self-attention layers as a viable and much more efficient alternative to RNNs. Borrowing from an Encoder-Decoder architecture (as was prevalent during the day) the researchers were able to get state-of-the-art performance on a multitude of Language modelling tasks. All subsequent models like BERT, GPT-2/GPT-3, and LLama-2 have been built on this foundational architecture.

## Types of Bias/Risks

**Historical Bias** - This kind of bias seeps into our ML models because of historical stereotypes. Large Language Models have been known to be plagued by such biases.

For example, a language model trained on the data from the 1960s assigns job roles in a very gender-biased way (i.e. it has the propensity to assign women more to secretary roles in comparison to men). In a similar vein, the language model learns to be biased against certain disadvantaged minority groups which have historically been on the sidelines of society.

**Representation Bias** - This is a classic example of sampling bias when the data used for training is not representative of the entire population. Take the example of facial recognition systems, if the system has been trained on faces of a particular community under a specific lighting condition, then it fails to recognise faces in the wild. In order for our AI models to be more inclusive, we need to collect data from a diverse population which is representative of the ground truth.

**Measurement Bias** - Measurement bias happens when we do not select the right features for the problem at hand or we choose proxy features/proxy class labels which are not representative of the desired outcome. It might also be inherent to the way we collect the data (i.e. collecting seasonal data for example during an online sale and using it to make recommendations for the entire year).

**Perpetuating Echo Chambers and Filter Bubbles** - When an LLM agent interacts with a user, depending on the user's feedback it can overfit the user's confirmation bias. This leads to the creation of filter bubbles and echo chambers where the user is shown content or provided recommendations which are aligned with their worldview no matter how distorted that might be.

**Adversarial Attacks** - Adversarial attacks on LLMs come in various forms but they have the common underlying of prompt modification in the form of - Prompt Injection, Prompt Leaking and Jailbreak Prompt. These techniques are used to reveal the inner workings of the model, reveal contents of the prompt and even trick the model into producing harmful content for the user. Certain models have guardrails built into them, however, it is non-trivial to come up with a foolproof solution.

**Toxic and Hateful Content** - We are all well-versed in how generative AI is used at a large scale to generate fake news which propagates toxic, hateful and xenophobic content. Although most platforms have some guardrails to deter toxic propaganda, these are far from enough. We have seen implications of these during elections and political upheavals across the world.

**Pervasive Nature of Deep Fakes** - Deep fakes which started out as something innocuous has taken a very sinister turn these days. They have been used to seek revenge from political entities and defame celebrities. What's even more concerning is that they are becoming indistinguishable from organic content. This sets up a cat-and-mouse game where the actors are trying to beat each other.

## Mitigation Mechanisms

1. **Increase Representation of the Dataset** - The web-scale datasets that are used to train ML models have an inherent bias against them which seeps into the model predictions. These accentuate and perpetuate the existing biases in the society. In order to mitigate these, we need more representational data which paints a clearer picture of the real world. Additionally, we need to comprehend which are the features on which the model is getting biased on so that they are removed or downgraded.
2. **Use Human in the Loop Feedback** - Setting up human guardrails to LLM responses and even the data used for training LLMs has recently gained much traction. Frameworks like RLHF (Reinforcement Learning Human Feedback) seek human feedback on the outputs of the LLMs. Open AI has also recently started to use human annotators to create training data for its models.
3. **Setup Model Audits and Red Teaming** - No model is immune to adversarial attacks, it would be prescient to be aware of these limitations even before it is deployed to production. Red teaming is a popular process used in cybersecurity which can be adapted to the Machine Learning use case in order to reveal inconsistencies and biases in the model predictions. Auditing the model to ensure it not only meets the defined SLAs but is also in compliance with the governmental regulations is a good practice especially now that many governmental agencies are making it mandatory to have such checks and balances in place.
4. **Modify the evaluation criteria for the models** - Objective functions to evaluate model performance only optimize the statistical measures of Precision, Recall and other salient metrics. However, we also need to focus on other measures which de-bias our models, make their predictions more fair and diverse. These can be incorporated into the model evaluation function during the training phase. This will ensure that the model is not anchored on its training data but is able to adapt to the real-world scenarios as well.

## Key Takeaways

1. Generative AI is here to stay and has the potential to enhance and augment human capabilities.
2. The web has played a pivotal role in proliferating the adoption of Generative AI at all stages of its research and development process
3. We currently lack best practices and a formal framework for defining the ethical implications of this technology there are early regulations around this (i.e. EU AI Act, the US Congress AI Bill etc.)

4. Since the landscape of Generative AI is constantly evolving, it's difficult to predict where this will end up, so the sooner we start having discussions on the future risks and potential mitigation strategies the better.
5. We need a diverse and interdisciplinary group of people to join the efforts around developing this ethical framework, these include - social scientists, economists, philosophers, policy experts, machine learning engineers etc.

## Conclusion

As a part of this document, we have highlighted the various risks and mitigation mechanisms which face LLMs today. We have also briefly touched upon the history and evolution of the models. Finally, we have addressed certain pertinent questions regarding the future that we envision for incorporating ethics as a part of the developers' toolkit. We are certain that Generative AI is here to stay and its benefits outweigh the potential risks. We as a society need to ensure that we have policies and practices in place to reap the benefits of this technology and minimize the harm.

## References

- [1] <https://aeon.co/essays/can-philosophy-help-us-get-a-grip-on-the-consequences-of-ai>
- [2] <https://www.vox.com/future-perfect/2024/2/28/24083814/google-gemini-ai-bias-ethics>
- [3] <https://www.forbes.com/sites/forbestechcouncil/2023/09/06/navigating-the-biases-in-llm-generative-ai-a-guide-to-responsible-implementation/?sh=33e47fbf5cd2>
- [4] <https://www2.deloitte.com/content/dam/Deloitte/us/Documents/consulting/us-ai-institute-generative-ai-and-the-future-of-work.pdf>
- [5] <https://www.bloomberg.com/graphics/2023-generative-ai-bias/?embedded-checkout=true>
- [6] <https://theconversation.com/eliminating-bias-in-ai-may-be-impossible-a-computer-scientist-explains-how-to-tame-it-instead-208611>