PHILOSOPHY / AI ETHICS / 01

Undergraduate Dissertation: Social Biases in Generative Artificial Intelligence

An analysis into what exactly gender bias in AI is, and practical solutions to redress it.

Davey Jackson

Abstract

This dissertation is a bipartite examination into the nature of Gender Bias in Artificial Intelligence, both philosophical and practical. The first part consists of asking what we mean philosophically by wanting “unbiased” Artificial Intelligence, and in an argumentative essay attempts to motivate this intuitive notion of “unbiased” as wanting the AI to represent an Ideal World, and then defining this Ideal World model as a concrete and well defined model as a target for AI that we want to be “unbiased”. The second half consists of an overview of existing techniques in AI to modify an AI system to better align with human expectations, specifically in data manipulation, prompt manipulation and fine-tuning, and an analysis of their efficacy in implementing our Ideal World model, with a conclusion that the initial two techniques are not effective, but using RLHF or similar fine tuning would allow an effective implementation of the Ideal World model in production AI systems.

Contents

Abstract 1

Preface 3

Research Questions and Essay Structure 3

Background 4

Bias in Pre-Generative AI 4

Generative AI 5

Analysis of Bias 5

The alignment problem 5

Should we be concerned? 7

The Ideal World 7

Balancing Interests 9

Application in Practice 9

Generative AI Background 9

The two problems 10

Dataset Manipulation 11

Prompt Manipulation 13

Fine tuning 15

Practical Application Conclusion 17

Conclusion 18

Bibliography 19

Preface

When you ask a Generative Artificial Intelligence (Gen AI) to generate you a picture of an executive, the system (assuming no specific adjustments to remedy this issue) would be far more likely to generate a picture of a man than a woman. This is very clearly an issue, the AI system is biased against women.

However, when you ask that Gen AI system to generate you a picture of a mixed gender boardroom the AI will draw the men wearing men's officewear and the women wearing women's officewear. This is very clearly desired behaviour, it would not even cross most people's minds to consider the moral ramifications and if the AI were to generate men wearing skirts it would be considered faulty by almost all users.

When discussing the problem of gender bias in artificial intelligence it is very easy to say, and very commonly said by those creating maxims for how one might create ethical AI (Lambert, 2021) (Future of Life Institute, 2017), that we should avoid gender bias. However, by looking at the two examples above we can very clearly see that the question of gender bias is very much not one that is cut and dry. For a Generative AI system to function effectively represent the world (or it's dataset) it must necessarily encode some information about the implicit differences between genders we see expressed in the real world, yet clearly some, but not all, of those differences are what we consider to be bias and should not be represented.

To add further to the issue, modern Generative AI systems are extremely complex beasts whose inner workings are arcane even to researchers in the field and rely on massive datasets that could not be feasibly read or seen by a single human to verify them. Making such complex systems meet our actual desires, what is known as an alignment problem in AI, is not a simple task. To create an AI system that is unbiased in gender, even with a perfect definition of bias, requires using techniques such as fine tuning, pretraining, RLHF, and other such tools, to try and nudge the AI system to move closer to our aligned goals.

Research Questions and Essay Structure

With the fundamental issues at hand raised, it would be useful to formalise them into the two core research questions that this essay will try and answer:

1. What is an effective framework to understand the problem of gender bias in Generative Artificial Intelligence
2. How can we apply such a framework using practical training techniques to rectify the problem in actual Generative AI systems

The essay will consist of two sections, each trying to address one of the two questions. The first section is going to be an argumentative essay tackling the moral philosophy of gender bias in AI, and the second section will consist of a summary and analysis of existing techniques in Generative AI with respect to the conclusions drawn within the first section.

The aim with our philosophical analysis is not just to create a hypothetical idea of what gender bias might be but rather use an interdisciplinary understanding of the practical requirements of AI training to build a framework that is both sound philosophically but can also be applied to real world AI models. And similarly, the analysis of AI training techniques will not simply be a technical rundown of the modern state-of-the-art techniques but rely on an understanding of the underlying philosophical principles to inform an evaluation about what techniques work the best.

Background

Bias in Pre-Generative AI

Gender Bias in Artificial Intelligence has been a significant issue since essentially AI systems have been asked to interact with the real world. Perhaps one of the most famous examples is that of an attempted AI based CV screening tool that Amazon attempted to develop in 2014, and which the project was eventually scrapped in 2017 due to an inability to avoid bias (Dastin, 2018). The system was trained on historic examples of people who had been hired by Amazon as examples of what a good CV would be, however because Amazon had a significant bias in existing hiring practices, with a predominantly male workforce, the AI system was trained to associate indicators of a male CV with being a successful candidate.

The Amazon case is a particularly useful example as it demonstrates a significant number of the fundamental problems that exist with the intersection of gender, or really any existing social reality, and AI. The system was not built to be sexist, the AI was not given any information on the gender of the applicant and there was no intention for it to discriminate as such, however the issue was the dataset. Because there existed already existing biases within the dataset the were using, in this case the hiring practices of Amazon, the AI system ended up learning these implicit biases and using them in its recommendation. This is fundamentally an alignment problem, as will be explained later in this essay.

In 2016 an AI system used by the US to judge likelihood of recidivism was criticised for being biased against black people by ProPublica (Angwin *et al.*, 2016), using the relative statistical likelihood of black people being given a certain score. However, in a paper by Chouldechova (Chouldechova, 2017) responding to the article they argued that such outcomes would be apparent in a system that was free from predictive bias, but rather we would see a disparate over scoring of black people simply due to the way the maths works out. In addition to being another example of the long relevance of social biases (in this case race) within AI systems, it presents a good example of the moral quandaries that we face. A system that is unbiased, in a mathematical sense, will tend to be harsher on average to minorities that are statistically more likely to reoffend, beyond the simple increase in likelihood due to them belonging to that group. However if we try and adjust the system to create similar rates of erroneous assessment of higher recidivism, we must necessarily bias the base results and reduce the likelihood of accurate high scores for the minority. Even when we can perform statistical tests for bias we still must make a moral judgment about what kind of bias we want, let alone with Generative AI systems that are far less cut and dry.

Generative AI

In the past few years the field of Generative Artificial Intelligence has massively exploded, with Large Language Models (LLMs) like ChatGPT and Diffusion based systems like Midjourney becoming staples in everyday life and the industry becoming one of the largest in the world. The technical details of how such systems work will be left for the later in this essay, but what is relevant now is that they have not escaped the issues of being biased just like earlier classification based systems.

A 2024 study by the International Research Centre on Artificial Intelligence (UNESCO and IRCAI, 2024) found systemic biases against women in LLMs. When asking them to complete sentences such as the "the man was" and "the woman was", there were significantly more negative responses for women than men in raw LLMs tested (ie those that hadn't undergone fine tuning through human feedback). However ChatGPT, which uses RLHF to tune the output from the raw language model, did significantly better.

We see similar issues with image generation AI tools, with Dalle 2 prior to mitigation work by Open AI creating images of almost exclusively white males for prompts like firefighter or CEO, and white women for the prompt of teacher (OpenAI, 2023).

There has been widespread condemnation of these issues, with the UN's Framework for Ethical AI as an example stating that AI should be non-biased, and genderless (Lambert, 2021), but it does not state exactly what they mean by that. More broadly the field has generally been quite ad hoc in their philosophical approach to addressing bias, when obvious issues of bias arise like racist or sexist prompts then fixes are applied, ranging from well developed internal changes to the system to applying "diverse" appellations to the prompts to create less homogenous inputs.

Analysis of Bias

The alignment problem

To start talking about the problem of gender bias in Artificial Intelligence, first we must cover the idea of alignment. In AI alignment refers to how well the aims (and outcomes) of the AI system is aligned with the intended aims of the human authors of the system. Alignment problems represent the core of many problems in the wider field of AI safety, as defining the exact parameters that humans want in a way that the system understands is extremely tricky.

To go back to the earlier example of the Amazon CV screening system, in that case the intended aim of the authors was to create a system that was able to classify if a CV is good or not. To do that they took past CVs of successful applicants as examples of good CVs and trained the AI to match to those. The AI was therefore not attempting to judge if a CV was good, but rather simply how similar it was to past successful CVs, including elements like gendered language that were not necessarily relevant to the quality of the CV, but very relevant to the similarity of the CV to past ones, creating an alignment problem.

In the case of Generative AI, with modern systems they are all fundamentally being trained to replicate the patterns they find in the massive datasets on which they are trained. A language model will simply take in the prompt and attempt to add the next word in based on how it has seen sentences go on in it's dataset, and image generator will try and create an image that matches the prompt based on the images in the dataset. This reflectivity of the dataset is the fundamental aim of any current Generative AI system.

Because these datasets are massive and contain pretty much the sum total of all the images or text an AI firm can procure, they are almost universally reflective of the broader social biases in the society that created the data used. There are simply far more images of white males in suits in offices than there are of women or people of colour, there are far more sentences written in the English corpus that are derogatory towards women that there are derogatory towards men, and these AI systems will reflect that.

That is where the alignment problem comes in, because it is clear that the authors do not generally want these kinds of inherent social biases to be reflected in the output of their AI models, and so they are misaligned. However, the issue is that we clearly do want them to be broadly reflective of the dataset, that is their entire purpose.

The first response to this might be to suggest that we would like the AI systems to be reflective of the physical realities of our world but not reflect the implicit social biases of our world, but this does not really stand up to scrutiny. As was said in the example in the introduction, we expect that the output of the AI models when generating men or women should generate them in their respective gendered clothing, it is clear that we want the AI to encode some kind of reflection of gendered social norms.

Instead, I posit that when these notions of AI that are "unbiased" are put forward, what is really implicitly being suggested is that the AI is reflective of a "Ideal World". The idea of a system that is "unbiased" is one that is fairly laughable, because any system that creates an output is going to have some kind of bias somewhere. If we were to try and create some reasonable definition of what unbiased is thought, it feels to me like the only coherent notion would be that of a system that accurately reflects reality, or the given dataset, because that is simply the system giving output irrespective of the author's intent.

When people advocate for unbiased AI though they do not advocate for this neutral reflective model, they advocate that for any particular job it should be a 50/50 split male/female, even if that is far from reality, or that there is always an equal mix of various racial groups in the images. What they're advocating for is a model that generates images as if there was no gender or racial bias in hiring for any particular job or being in any particular circumstance, that is what they mean by unbiased, hence my assertation of an "Ideal World".

With an understanding of the alignment problem, and the idea of the "Ideal World", we can now deconstruct this notion of the Ideal World in order to build our framework of what we consider "gender bias".

Should we be concerned?

Before breaking down our Ideal World notion, it is first worth asking if we should be concerned about gender bias in Gen AI in the first place, as we've been taking the notion that it is bad at face value until now. The idea that we should simply have our datasets create a raw model of the world and then let them loose, and what will be will be, is certainly one you can find espoused on various internet fora by machine learning experts who certainly did not sign up for any of this ethical nonsense. There is a certain raw simplicity in simply absolving the AI models of any responsibility and placing the blame on society at large or the people using these models, however by virtue of writing this essay it is clear I do not agree with it.

There is of course the equally simple argument that sexism (and other forms of bigotry) are simply bad, ontologically. If you have a computer program that is biased towards saying sexist remarks, then it is a net negative unto this world because it is sexist and releasing such a program and not attempting to fix it is a moral wrong because you are spreading sexism. Such an argument I feel is fairly strong, however it places a lot of moral agency into the hands of a program that can be seen as simply a tool. An image generation program does not do anything on its own, it simply creates images from the prompts you give it, if some prompts generate sexist output then simply use other prompts.

A more robust argument against biased AI systems is one of reinforcement of social biases. Whilst these AI systems simply learn biases from society at large in their training, if these biases continue into production with models that see widespread usage then they are not only reflecting biases but actively reinforcing them. As an example, many AI image generation tools are used as filters on online apps that are often used widely by young teens. If a teenage girl uses an AI filter to see what she would look like as a pirate (as an example) and it produces a sexualised image in her likeness, whereas it simply produces a fairly standard pirate in their likeness for her male friends, that is going to reinforce objectifying notions in that girl, and probably all those around her. That's just a single fairly straight cut example, but LLMs are used constantly to write text for all sorts of applications across the world, even if only implicitly if they are reflecting social biases in their text that is going to have a massive reinforcing effect on those biases in society at large.

As such I do think we can firmly say that these widely propagated and widely used models have a duty to at least try and mitigate the social biases present in their outputs, or they are to some degree morally responsible for the continued existence of these biases in our society.

The Ideal World

With that brief aside completed, we can now move on to discussing the Ideal World. The Ideal World is one that is **as similar to our world as possible but lacking any kind of social prejudices that are not considered relevant to the generated work**. This statement contains a significant number of key criteria that I will expand on, but it stands as the core definition of the Ideal World model that we will consider the target for Generative AI.

The first key element is that it is as similar to our world as possible, the Ideal World is not meant to represent some kind of fictional utopia but rather an extremely close idealisation of our current

reality. The idea of “our world” is also a slight bit of fiction, as Generative AI can and does create representations of fictional worlds. In such cases, the aim is very much similar, if one asks an AI to create a story set in a space faring civilisation then the Ideal World in that case is simply as similar to this fictional world as possible but still meeting the other criteria.

The second element, and crux of the whole definition, is that it is lacking any of the social prejudices that we face in our world. Now when parsing this it is important to note that what we’re describing, our Ideal World, is not an output for the AI system, but rather meant to be representative of the world that the AI is learning through its training process. As such when the AI creates an image of a workplace, one would expect there to be an even distribution of sexes, if the AI was to write an email to a female coworker or a male coworker it would use the same tone of voice, as the world it would have learned from would not use a different tone of voice.

In a way this is just essentially moving the heavy lifting from the word bias into the word prejudice, but by reframing the question the line becomes significantly easier to draw. We do not generally consider it prejudicial that there is an expected norm for what clothes individuals wear, but a firm choosing to hire men over women would be considered to be prejudicial. There would definitely exist feminist scholars who disagree with this assertion, and that any examples of binary gender would certainly not be an Ideal World, but in practice most people using an AI model in their day to day lives are perfectly happy to live in a world with some level of gender norms. When we discuss prejudice here there has to be some appeal to common sensibility, if the common person would not take offense at it in principle, then it can hardly be considered prejudicial.

The “in principle” there is key, the average person might not take umbridge with any particular piece of art or writing that has a woman sexualised, but it would be fair to say that the broad phenomenon of sexualisation of women in media above men is considered prejudicial, and so the world that this AI is trained on would have media that is relatively even. That is not to say there would be no sexualised content, but rather that there would not be an implicit assumption that if it is a woman it is sexualised by default.

This test of prejudice also allows us to account for a broad spectrum of issues in terms of representations that are otherwise fairly hard to tackle. For example, if one were to generate an image of a couple cuddling in bed, it would be likely that most representations are going to be of the man being the big spoon, which is definitely an learned social bias. However, it is unlikely that any reasonable person would consider that to be prejudicial and so it may stay.

The last criterion that we have is that of relevancy. That is to say that if it is necessary for the representation created by the Gen AI to be based in a world that has some kind of prejudice then it should. This is not to cover things like prompting for an office full of white males in suits, which is entirely within the remit of the model as is, as under our model it is the world the dataset represents that we are targeting, not the output. But if one asked for an image of a Nazi solider in WW2 it would be unreasonable to assume a world in which the Nazis lacked any kind of prejudice. Similarly, if one were to ask for historic photos of soldiers in general, whilst there was

prejudice in not allowing women to serve in the military it would be considered a significant fault of a Gen AI model to create images of women serving in the trenches of WW1 unprompted.

Unfortunately, this is where the criteria start to get slightly muddy. For example, given the very, very significant lack of electricians that exist in the current day and the existence of Women's soldiery in the USSR during WW2, based simply of statistical chance it would be hard to say why showing pictures of women electricians as often as men is a part of our system, but showing pictures of women as WW2 soldiers would not be. Again, reliance on a test of reasonableness seems to be the primary factor, would an intelligent user upon seeing a generated piece of content deem it unreasonable off of the prompt. One would not consider a woman electrician unreasonable, but a woman WW2 solider (off simply the prompt WW2 soldier) likely would be. To be slightly more objective, if the category of the prompt necessarily implies some level of prejudice (ie the Nazis) or if the exceptions to the category are so rare and very specific as to not at all be representative of the broader asked for class (ie Soviet Soldiers) then it would be considered unreasonable to suspend prejudice under the Ideal World model.

Balancing Interests

The fundamental purpose of Generative AI is to create a useful output. If it does not do that then there is fundamentally no reason for it to exist. As was discussed earlier, there is also a moral obligation on the part of the AI authors to create an AI that is not propagating ideas that are bigoted and further reinforcing social biases.

It is the balancing of these two competing factors that the Ideal World model is aimed to addressing. By focusing on specifically representing a world that lacks outright prejudice, but not one that lacks social biases altogether, it creates a target world for an AI model that is able to produce useful outputs whilst also mitigating the social harms of a biased AI models.

Application in Practice

Generative AI Background

Now that we have our ideal world model, the challenge is to try and use it to improve actual production Gen AI systems. To begin with some background information on the workings of Gen AI.

To discuss all possible forms of Gen AI would be significantly outside the scope of this essay, but there are fundamentally two architectures that have dominated the scene in the last 4 years, and which are relevant to our discussion: the transformer and the diffusion model.

Transformers have been the model behind the boom in LLMs, with unparalleled ability to train on massive corpuses of text using parallel computing. A text generation transformer is trained by giving it massive amounts of text and having it learn the associative relationship between the words in the text based on their meanings and the context around them.

Diffusion models also rely on large, unlabelled datasets, but work by essentially being trained to denoise a noisy image and then being trained to denoise entirely random white noise, creating an entirely new image based only the prompts that they are provided.

What is key for both of these models is that they are very complex internally, and so cannot be easily modified or tuned, the way one might try and fix a linear regression model by adjusting one of the parameters. In a large AI model, there might be on the order of billions or even trillions of parameters (Schreiner, 2023). As a result, in order to modify these large, unspecific models, we need to use various filtering, pretraining, fine tuning and similar techniques, all of which are useful in their own way, but limited in others.

The two problems

As mentioned above, AI systems are reflective of the dataset that we provide them, and the biases that they inherit are biases that are reflected in the dataset. The quality of modern Gen AI systems has come about largely because they use extremely large datasets, the significant innovation of OpenAI's GPT series of transformers was the size of dataset they used more so than anything architectural. Because these datasets are so large the biases they have are the biases of societies that they come from, they are broadly reflective of the entire corpus of text or of images that a particular society has created. They do also inherit some bias from who can create a lot of data, images of white people are far more common in these large datasets because most images available on the web are from predominantly white western countries who have had broad access to digital cameras for decades now.

When discussing the existence of these biases we can broadly classify the problem into two distinct categories, for which some solutions may be more practical than others. The first problem is distributional, that the distribution of data points is skewed in a particular direction which causes the output to be skewed similarly, which creates a biased output. This is the predominant issue facing image generation AI systems, there are far more images of males as doctors and lawyers and far more of women as teacher or nurses, or there are significantly more images of women sexualised than not in particular contexts, and so outputs will default to being sexualised because that is the content they receive. But we also see this in text generation, when asked to complete a sentence saying "women are" there are far more examples of sexist content than not in the dataset (UNESCO and IRCAI, 2024), and so the sentence is completed in a sexist fashion.

The second problem is integral, that there are implicit biases encoded within the data that the AI is using to create the dataset. When women are talking in the vast majority of texts in a corpus they are likely to be presented as deferent to their male superiors, and so when the AI is asked to generate text of a woman speaking to a male superior it will default to that tone, or similarly in images of business meetings it is likely that the person presiding will be a man. These are also representations of distribution of data in the dataset, but the distinguishing factor is that it is not possible to simply disentangle such representations from the wider data they are within, compared to the distributional problems where the issue is the quantity of discrete specific examples.

These two problems are inextricably linked, and many of the solutions will tackle both, but there are some which may only work for one or the other. Additionally, as mentioned, because image generation relies on very independent discrete objects, the images, versus the far larger and more complex datapoints of texts in a corpus, distributional problems tend to appear more often in image generation and integral problems in text data sets.

Dataset Manipulation

Data Filtration

The first technique worth discussing is that of filtering the datasets used. Given the reflective nature of AI systems, the first technique to create an AI that is reflective of our Ideal World model is to try and filter the dataset in order an Ideal World dataset that the AI can then reflect, but unfortunately such a technique does not work for our purposes, on both practical grounds, but additionally filtering a dataset fundamentally cannot create an Ideal World AI system.

Specifically when we talk about data filtration, we are talking about taking the entire raw dataset before the training of the AI model and going through the dataset in order to remove content that is going to create biases in the model the demonstrate prejudice or alter the distribution of the model in order to better create a distribution that we like.

Such systems do exist in practice, for example OpenAI filtered out content that contained violent or sexual imagery in the creation of Dalle 3 (OpenAI, 2023), and similar examples exist in text, for example FalconLLM similarly removed sexual content from their dataset (*Falcon 3*, 2024). Such filtering is significantly easier said than done though, because at the size of these datasets you cannot manually filter the dataset, and as a result you must rely on using other AI tools in order to filter the dataset and create the data to be used in the final AI.

Such tools are effective when looking at (relatively) easy to define discrete categories, such as sexual content, but even then, there are issues with AI classifiers detecting simple images of women in swimwear as sexual, versus not for men. Trying to generate AI systems that can identify if content is somehow implicitly biased is significantly harder, essentially trying to solve the entire problem being posited in this essay. That alone likely makes data filtration unsuitable for our purposes.

The filtration is also likely to have an impact on the quality and distribution of the output product. When Open AI filtered out images of sexual content that reduced the amount of women in their dataset, which skewed the outcomes (OpenAI, 2023), but additionally by removing any data from the dataset you are harming the overall final product. If one was to try and remove enough images of men as doctors to create parity with women doctors, that might necessitate removing enough images that the AI is bad at creating images of doctors altogether, if one was to try and do such filtration across all professions it might make the AI simply bad at generating people, altogether. If one is entirely removing a category, like sexual images, it might be reasonable to see the AI degrade in quality in that sector, but for distributional changes to the dataset that degradation in quality is unacceptable. The end goal is an AI that can effectively represent the Ideal World, and whilst bias is important, the ability to effectively represent the world as it is is also necessary.

The fundamental issue

The above factors already disqualify data filtration is a method to create an Ideal World AI, but there is another, significantly more fundamental issue, which makes such a system untenable for our aims. An Ideal World AI, as we have posited it, wants to create a world that is as close to ours as possible, but also one that can represent content that might otherwise be objectionable if it is necessary, not simply an AI that has no notion of any kind of objectionable content whatsoever.

Removing content that has a gender bias would simply result in an AI that could not conceive of the notion, if one asked it to create a story set in a 1950s nuclear family it would be farcical without an understanding of the strict gender norms governing how the members of the family act.

In order to achieve our Ideal World AI, what is necessary is an AI system that has learnt and understands the implicit biases that exist in the world as it is, but an AI system that then somehow learns that said biases are bad, and as such the output of the AI needs to be altered to account for the biases in the training set. A data filtration system simply cannot accomplish such an outcome because it does not learn the bad thing in the first place.

This indicates that what is necessary is some kind of fine-tuning system, which is a system used after the initial training to “fine tune” the AI in order to create output that is better aligned to the aims of the human users. The class of fine-tuning algorithms is large, and we will explore such systems in depth later, but first there are some other possibilities that should be explored.

Data Augmentation

The other possibility to mess with a dataset, outside of filtration, is augmentation. Data Augmentation is broad class of systems that seek to add more data to your dataset, often by manipulating data from your dataset in specific ways in order to make it more robust. For example, if one were trying to classify images your might slightly rotate, blur, or add static to the images to create new arrangements of pixels that an AI classifier would see as different, to make it more robust to those modifications in practice.

In our case data augmentation would involve identifying distributional issues in our dataset and then using an already existing generative AI in order to create data that would fix the distribution, for example by creating a significant number of images of women in various professions that are otherwise male dominated so that the total number of images in the dataset is relatively balanced.

This idea is not quite as fundamentally flawed as the idea of data filtration, but it does have some significant hurdles to overcome. The most obvious is that such a system is only effective for a very small class of specific distributional problems, where removing such a distribution from the dataset would not cause issues and where it is very discrete and easy to identify. It also requires manual intervention, as there is no way to somehow classify every distributional issue in the dataset and create new pictures, for each identified problem someone must instruct an AI to generate the images.

The bigger issue though is that artificial data is very problematic. When we feed the AI actual images of the world, it learns to try and represent the world, if we feed it AI images, we are creating an AI that learns to generate reflections of another AI image generator. Any artificial data has this issue, unless it is perfectly reflective of the dataset you actually want to learn about it will cause a degradation in quality as the AI instead aligns to your artificial data. Current AI systems are definitely not perfectly reflective, AI images or AI text can be fairly regularly clocked by people because of subtle, but extant, idiosyncrasies.

Given the relatively limited scope of the solution and the significant issues in quality it could create, especially as correcting such distributions might take up a significant proportion of the existing data, such data augmentation is probably not an optimal path to take. However, it is relatively simple, and so for a smaller AI system that wants to be more aligned but cannot afford the cost of some other solutions we will present later, augmenting the dataset with artificial data to adjust distributions might represent a viable option.

Prompt Manipulation

The next broad class of methods that we must cover, before going into fine-tuning, is prompt manipulation. These include things like prompt injection in image generation or the initial instructions to LLMs and are almost the polar opposite of dataset manipulation. Rather than seeking to modify the actual internal model of the AI, prompt manipulation simply alters what inputs we give the AI in order to get better aligned outputs. This can have mixed results, but is a solution very commonly used in practice so is worth exploring.

Prompt Injection

The first example of prompt manipulation we are going to look at is prompt injection. This is a technique used in image generation in order to try and counter the distributional issues that image generation systems face, where additional words such as “diverse” or “woman” are appended to the end of the prompt provided by the user before it is given to the AI, which then uses those prompts to alter the outputs (Sparkes, 2022).

This approach first saw use, in a very ham-fisted manner, after the initial explosion of AI image generators and subsequent investigations into their output biases, which created significant amounts of bad PR for the large tech companies who had released these models. In this early approach they simply appended “diverse” keywords onto the end of prompts provided by the AI with limited rhyme or reason (Ananya, 2024). That created issues such as AI generating images of black or brown people in Nazi or Confederate groups or simply refusing to generate white people when specifically asked to do so. These early implementations significantly degraded the quality of the output and were rapidly replaced.

The issues with this naïve approach to prompt injection should be fairly obvious, it is unable to generate large classes of images that aren’t diverse, it significantly takes away user agency and makes the system feel out of control, it can create odd artefacts and broadly is quite dumb. It also can only resolve the simplest surface level issues, providing little recourse to the problem whilst making the AI system a lot worse, such prompt injectors are simply bad.

Prompt injection is still used today, it is a significant part of Dalle 3's systems for ensuring that the images created are not biased, but the implementations are significantly more advanced (OpenAI, 2023). Perhaps the biggest difference is that Dalle 3 is exclusively used through Open AI's ChatGPT interface, and so the user is interacting with an LLM. The prompts are so much blindly injected, but rather the LLM will take the prompt and transform it into a prompt for the image engine, often not using the verbatim words the user inputted.

This hybridised approach is significantly more effective than the naïve approach, using an LLM to filter the input and then using the filtered input to generate an image. LLMs are able to read the text, and broader context of the conversation around it, and make a judgement on if appending such keywords is necessary and useful or if it would hard the image in some capacity. They can also broadly understand the prompts meaning and so react to varied prompts unlike the original systems that relied on specific keywords or similar basic triggers to add their diverse injections.

Such a system also does not lack any merits in efficacy, image generation AI systems are very responsive to prompts generally, which when unprompted will default to the social biases that underly the system. By overriding those biases, it is very easy to create outputs that do not represent said biases, which might otherwise be extremely hard to remove from the underlying AI system.

However, even with an LLM layer, there is still some loss in user agency and ability to create anything they want, and the system is still limited in solving issues that can be explicitly enumerated and addressed by adding on a keyword, not more subtle implicit biases. With that being said, if one were to try and create an image generation AI that was to use the Ideal World model, the overall efficacy of the system and ease of implantation with a good LLM mean it would be a useful step to add. In an ideal world we would not need to use such a system, the underlying AI would be representing a world without prejudice but given that perfection is unlikely to be seen in production systems, an additional layer that helps is not unwelcome.

System Prompts

The analogous concept in LLMs is the system prompt, the initial prompt given to the LLM before they interact with the user. Such prompts are ubiquitous, as the fine-tuning they receive involves significant training to react to said prompts, and for systems being used in commercial applications they are key to giving the AI the context of what they are supposed to do.

There is something appealing about the notion that we could simply tell the AI to not be sexist, and then it would not be sexist, that feels like how one would resolve such issues with a genuinely intelligent AI, but alas LLMs are not that. Using the system prompt for safety concerns has a lot of issues, it is easily possible to bypass or break the AI's instructions by trying to convince it otherwise or it may simply bypass them itself if the AI is not very well aligned.

However more importantly a system prompt is effective at initiating specific conversations or avoiding specific defined areas, but implicit social biases are not that. An LLM is not a thinking machine, it is simply generating the statistically most probable next word, and when given a simple instruction on not being sexist those few words are not likely to perfectly align it compared

to the probability distribution of all the other words in the context window. Relying on a specific written instruction simply is not viable in current AI.

There are currently very sophisticated reasoning models that have long internal chains of thought and are able to solve complex problems (Wei *et al.*, 2023), it is distinctly possible that one could task such a model with analysing its own text and criticising it for social biases, I am unaware of any such attempts in the literature and their efficacy but even if such a system worked, it would be largely impractical. Chain of Thought models generate many, many tokens of thought for limited output, using such a system to mitigate social biases would require tremendous amounts of compute when such goals can be achieved elsewhere.

Whilst chain-of-thought represents an interesting possible avenue of research, broadly we can say that using the system prompt, or similar prompting setups, to try and create an LLM that represents an Ideal World is not practical.

Fine tuning

Finally, we get into the most promising, both in theory and practice, area of large AI modification for our purposes. Fine-tuning is a very broad category of techniques, all of which involve taking an AI that has already been trained on a large body of unorganised data, to create it's own basic understanding of the domain, and then using domain specific training to tune that raw AI to be a useful tool for a specific task.

What makes fine tuning different to prompt manipulation is that prompt manipulation manipulates the input into the AI model, which is a fixed entity, whereas with fine tuning we are actively manipulating the AI in order to try and change the internal understanding that it has.

There are many ways to do such fine-tuning, often involving using domain specific information to make it useful in a specific subfield. For our purposes we are going to primarily focus on reinforcement learning based approaches, which use reward mechanisms to train the AI.

Reinforcement Learning with Human Feedback

As was enumerated a few times in the criticisms of the above methods, one of their core issues is that they can only tackle relatively simple issues of bias that can be explicitly enumerated and defined, whereas many of the biases we would like not to be represented without need are not things one can well define or create an explicit rule for.

It is intuitive that this is an issue for a system that is computerised, and relying on an automated reward function, and as such requires some explicit definition of what is good or bad. Fortunately, this is a well understood problem, trying to align an AI with human preferences that are not easy to define, and as such Reinforcement Learning with Human Feedback (RLHF) was created as a solution.

To explain RLHF, let us first start with a simpler example, that of a robot trying jump as high as possible. To do this, we would use simple reinforcement learning, an architecture by which an AI agent tries to do a task, and gets a reward/punishment based on its success, and then uses that

feedback to learn for the next attempt. We can simply reward the robot based on jump height, and as the robot jumps it gets more rewards the higher it gets.

Once the robot can jump, we might try and train it on a more complex task, in this example a backflip. Unfortunately we cannot simply get a number for backflipness like height, and if we try and measure any specific details about the robot, say the angle of the back, it will likely simply try and "game the system" as the robot is simply trying to maximise it's rewards, whatever we say they are.

What would work would be having a human observe the robot, and then give it some kind of backflip score, we can all identify what a backflip is when we see one. Unfortunately to train an AI robot takes hundreds if not thousands of attempts, and a human manually scoring each jump would be infeasible.

Instead, the solution is to have a classifier AI, and that AI is trained on human feedback. One gives a human observer videos of the robot backflipping, and the human rates their quality. Then the classifier AI trains itself to score the robot based on the video to match the human scores, and the robot uses that classifier as the reward function. With such a method you can have a robot learn how to do such a nebulously defined thing as a backflip. That method is the fundamental architecture of Reinforcement Learning with Human Feedback.

When applied to Gen AI, RLHF is arguably the second most important technological leap we have experienced, since the implementation of the transformer. Open AI used RLHF to take their raw GPT 3.5 model, which was a simple autocompleting LLM, and transform it into a model that was able to converse with people in natural language, that modified model was ChatGPT (and its less well-known precursor, InstructGPT) (OpenAI, 2022) Open AI was able to get their LLM conversational through this method, demonstrating the power of RLHF in aligning an AI with human expectations.

What is notable is that in testing, we can see that ChatGPT is significantly less likely to generate responses that are objectionable to people than other LLMs which do not have any kind of RLHF system (UNESCO and IRCAI, 2024), and as such are just using the raw output of their dataset. This suggests that RLHF is in fact efficacious in creating an AI that is less likely to exhibit unwanted biases in its output.

We can then theorise that if a Gen AI system was trained by human responders who rated the output of the AI according to our notion of an ideal world, then the AI would be able to effectively converge on our Ideal World model, satisfying all our requirements.

The primary issue with RLHF is the HF portion, RLHF requires significant amounts of time and money spent on human respondents, often in third world countries being paid relatively little, which makes it almost entirely the domain of large AI companies able to pay for that kind of feedback. It's also possible that if the respondents are in a country who's notion of what is prejudicial does not align with the notion we are using, it could create a situation where the AI is aligned with their aims, but not the aims of those using it in western countries, and respondents

who are paid by the response and expected to work at a fast rate may not be ideal at judging subtle implicit biases.

These problems are notable, but certainly not insurmountable, they simply require good training and workplace conditions for the people training the AI classifier. The cost issue is certainly still a major barrier, but that is unfortunately hard to avoid.

Reinforcement Learning with AI Feedback

The potential next step to RLHF, RLAIF, is one in which an AI system generates responses which are then critiqued by an AI model according to the criteria set out and then trained on that. RLAIF is used by Claude in their Constitutional AI model, which aims to create safe AI without relying on extensive human feedback (Anthropic, 2023).

Such a model would generate responses, and then self-critiques the response according to a set of guidelines (the eponymous Constitution in Claude's case), revises the responses accordingly, and then is fine-tuned on the revised responses. This method is fairly similar to our idea of using an LLMs system prompt to meet our criteria, but with some key differences. Because it is in the training stage, one can afford to expend significant amounts of compute having the AI actively self-critique and use its own internal reasoning, which LLM reasoning models have shown to be highly effective. Obviously RLAIF is only as useful as the AI which does the underlying feedback, but if one has access to an already strong language model, especially one trained on reasoning, then the feedback should be quite strong.

Some human feedback is still necessary to start the system off, but one can easily imagine an AI fed with the first segment of this paper and asked to critique either its own outputs or the outputs of another model generating text or images and then using that to train itself into aligning its outputs with our Ideal World.

Practical Application Conclusion

As we can see, it is perhaps unsurprising that when trying to apply a human definable but hard to explicitly enumerate concept, the best way we can find to train an AI system to align with that concept is RLHF, or more advanced developments of a similar concept. Using some kind of prompt manipulation, even with RLHF, is also probably advised simply because it is relatively cheap and simple when an LLM is involved, but the raw underlying models should not be messed with as it compromises the entire efficacy of the Gen AI.

Conclusion

In this essay I aim to have done two things. Firstly to have taken the intuitive and ill defined, but oft invoked, idea of “unbiased” and shown how we can effectively ground it as our explicit idea of an AI using the Ideal World model, aiming to represent a world “as similar to our world as possible but lacking any kind of social prejudices that are not considered relevant to the generated work”. Secondly to have shown how across the many possibilities of manipulating AI models to align with our Ideal World, using an RLHF (or similar) fine-tuning technique represents the optimal method to implement an Ideal World model in practice.

Bibliography

Alfano, M., Abedin, E., Reimann, R., Ferreira, M. and Cheong, M. (2024). 'Now you see me, now you don't: an exploration of religious exnomination in DALL-E'. *Ethics and Information Technology*, 26 (2), p. 27. doi: 10.1007/s10676-024-09760-y.

Ananya. (2024). 'AI image generators often give racist and sexist results: can they be fixed?' *Nature*, 627 (8005), pp. 722-725. doi: 10.1038/d41586-024-00674-9.

Angwin, J., Larson, J., Mattu , S. and Kirchner, L. (2016). 'Machine Bias'. *ProPublica*, 23 May.

Anthropic. (2023). *Claude's Constitution*. Available at: https://www.anthropic.com/news/claudes-constitution (Accessed: 21 May 2025).

Chouldechova, A. (2017). 'Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments'. *Big Data*, 5 (2), pp. 153-163. doi: 10.1089/big.2016.0047.

Contributors, A. (2023). *How ChatGPT turned generative AI into an "anything tool"*. *Ars Technica*. Available at: https://arstechnica.com/ai/2023/08/how-chatgpt-turned-generative-ai-into-an-anything-tool/ (Accessed: 21 May 2025).

Dastin, J. (2018). 'Insight - Amazon scraps secret AI recruiting tool that showed bias against women'. *Reuters*, 11 October. Available at: https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/ (Accessed: 21 May 2025).

Falcon 3. (2024). Available at: https://falconllm.tii.ae/falcon3/index.html (Accessed: 21 May 2025).

Future of Life Institute. (2017). 'Asilomar AI Principles', 8 November. Available at: https://futureoflife.org/open-letter/ai-principles/ (Accessed: 21 May 2025).

Lambert, L. (2021). *A Framework for Ethical AI at the United Nations*. UN Office for Information and Communications Technology.

Markov, T., Zhang, C., Agarwal, S., Eloundou, T., Lee, T., Adler, S., Jiang, A. and Weng, L. (2023). 'A Holistic Approach to Undesired Content Detection in the Real World'. arXiv. doi: 10.48550/arXiv.2208.03274.

Mattu, S., Larson, J., Angwin, J. and Lauren , K. (2013). 'How We Analyzed the COMPAS Recidivism Algorithm'. *ProPublica*, 23 May. Available at: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm (Accessed: 21 May 2025).

Mauro, G. and Schellmann, H. (2023). '"There is no standard": investigation finds AI algorithms objectify women's bodies'. *The Guardian*, 8 February. Available at: https://www.theguardian.com/technology/2023/feb/08/biased-ai-algorithms-racy-women-bodies (Accessed: 21 May 2025).

Mehan, J. E. (2024). *Artificial Intelligence: Ethical, social, and security impacts for the present and the future, Second edition*. IT Governance Publishing. doi: 10.2307/jj.18137933.

Mu, T., Helyar, A., Heidecke, J., Achiam, J., Vallone, A., Kivlichan, I., Lin, M., Beutel, A., Schulman, J. and Weng, L. (2024). 'Rule Based Rewards for Language Model Safety'. arXiv. doi: 10.48550/arXiv.2411.01111.

OpenAI. (2022a). *Aligning language models to follow instructions*. Available at: https://openai.com/index/instruction-following/ (Accessed: 21 May 2025).

OpenAI. (2022b). *DALL·E 2 pre-training mitigations*. Available at: https://openai.com/index/dall-e-2-pre-training-mitigations/ (Accessed: 21 May 2025).

OpenAI. (2023). 'DALL·E 3 System Card'.

Schreiner, M. (2023). *GPT-4 architecture, datasets, costs and more leaked*. *THE DECODER*. Available at: https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/ (Accessed: 21 May 2025).

Sparkes, M. (2022). 'AI art tool covertly alters requests'. *New Scientist*, 255 (3397), p. 10. doi: 10.1016/S0262-4079(22)01329-X.

UNESCO and IRCAI. (2024). 'Challenging systematic prejudices: an investigation into bias against women and girls in large language models'. in.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. and Zhou, D. (2023). 'Chain-of-Thought Prompting Elicits Reasoning in Large Language Models'. arXiv. doi: 10.48550/arXiv.2201.11903.

YEUNG, D., KHAN, I., KALRA, N. and OSOBA, O. A. (2021). *Identifying Systemic Bias in the Acquisition of Machine Learning Decision Aids for Law Enforcement Applications*. RAND.