

A nuanced view of bias in language models

 viden.ai/en/a-nuanced-view-of-bias-in-language-models/

VIDEN.AI

   [Log in](#) [Subscribe](#)

The use of generative artificial intelligence, such as artificial intelligence, is a significant concern. ChatGPT is also becoming increasingly prevalent in education. When we incorporate new technology into our work and education, we must understand its possibilities, particularly its limitations. Large language models, such as ChatGPT, are powerful tools that offer excellent opportunities, but we must also be very critical of them. There are potentially a lot of biases "built-in" in the systems that we must acknowledge and be fully aware of when we use them - and mainly when we use them for teaching.

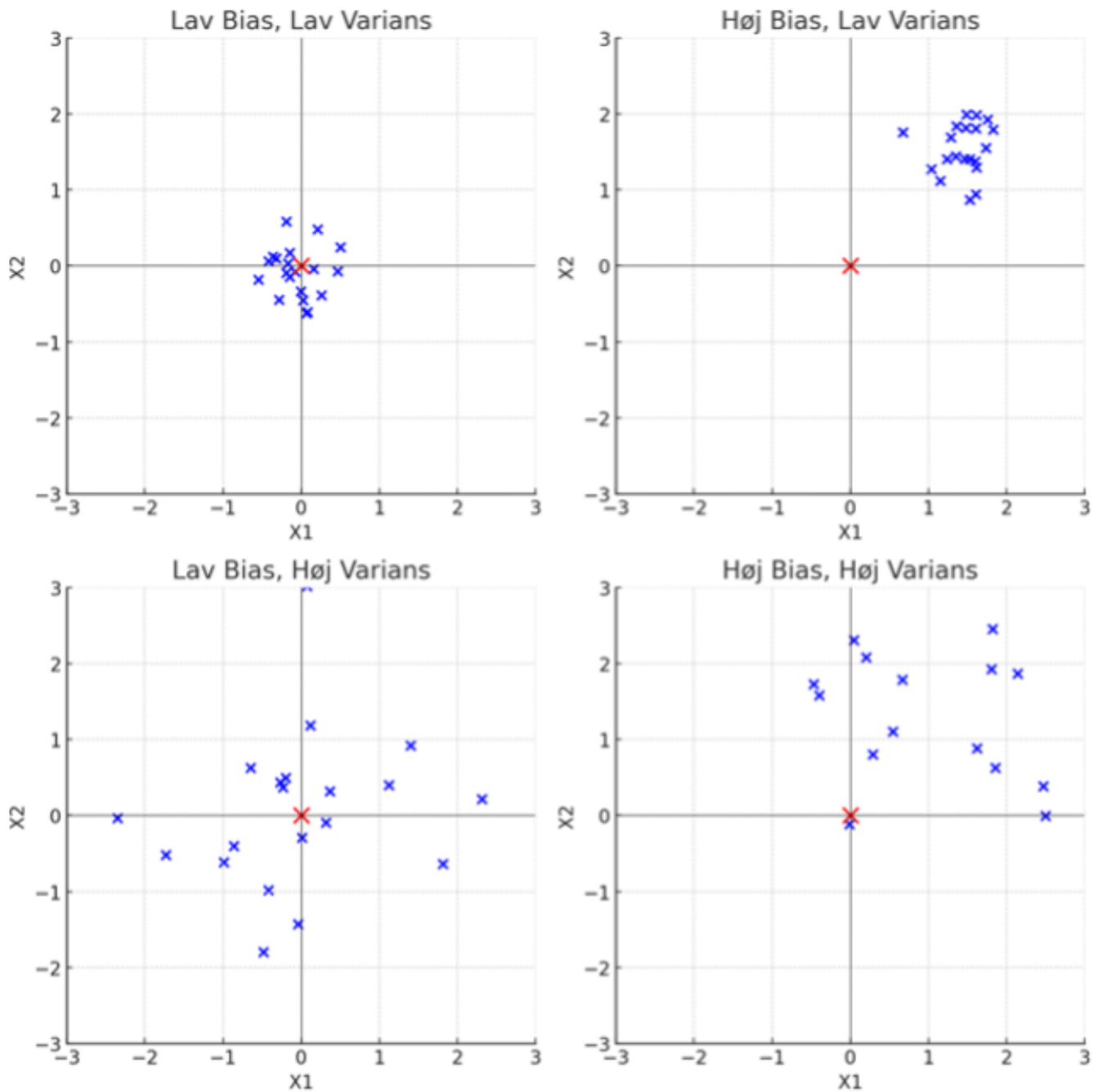
This article will focus on a wide range of possible biases in language models and is intended as a resource that can be used when teaching critical attitudes, source criticism, and responsible use of artificial intelligence in teaching.

The concept of bias

When we talk about bias, the term "bias" is often used – also in Danish. The word bias comes from English, but according to the Danish Language Council, it was used as a Danish word as early as 1953 and meant "bias, tendency or bias."

In contrast, high bias represents where all arrows are close together but well away from the center. The variance is about how much spread there is between the individual arrows. If all arrows in a series are close together, there is low variance, while high variance describes the situation where the arrows are scattered throughout the disc.

In the figure below, the red cross corresponds to the center of the disc (bullseye), and the blue crosses represent the position of the darts.



Visualization is done with "Advanced Data Analysis" in ChatGPT-4

If we have to translate this into language models, statistical bias corresponds to how accurately the model predicts a text *relative to what we would expect*. The variance corresponds to how consistently the model corresponds to the expected. A static bias says something about how good the model is at writing from the data it is based on and *not* how good it is about the world we live in!

When writing and talking about bias in language models like ChatGPT, it's usually about the *mismatch between what the model writes and what we would like it to write* (the ideal world in the eyes of the individual?). In reality, it's more about our expectations and our *expectations* about the data used to train the models.

In this article, therefore, I only use the word bias about the designations of the various biases that can arise because most of the concepts come from English. Still, I stick to the idea of "biases" elsewhere.

Different types of bias

There is a wide range of potential biases that can occur in and through the use of generative language models. Many of them are "built-in" into the models and have arisen in data selection and during their training and implementation. In contrast, other biases occur when we use language models.

I will cover several possible biases in this article and explain how they can occur.

Machine bias / algorithmic bias:

Algorithmic bias refers to the biases in the data used to train significant language models such as ChatGPT. As these models are trained on large anthropogenic (or perhaps better to say humanly selected) datasets (e.g., texts from the Internet), they tend to perpetuate the inequalities present in the datasets, thus perpetuating, for example, stereotypes and discrimination. If data contains biases, the models are likely to do the same. White Western men write most of the texts available in the world, which creates a distortion. If a particular group of people has consistently been portrayed negatively in training data, the model may exhibit a similar negative angle in its responses. Therefore, language models may contain biases about gender, sexuality, religion, race, culture, age, socioeconomics, geography, etc.

Source: <https://towardsdatascience.com/algorithm-bias-in-artificial-intelligence-needs-to-be-discussed-and-addressed-8d369d675a70>

Another problem with algorithmic bias is that the training can inadvertently reinforce existing biases. This is because the models can "consider" frequently occurring opinions or depictions as "normal" or "typical". Since the models do not understand context, this can generate responses that may seem tendentious, even if it was not intended.

Availability bias

Availability bias usually refers to the human tendency that we often base our decisions and judgments of information on the knowledge that is readily available or fresh in our memory. This may mean that more prominent or new events may have a disproportionate impact on our decision-making, although they may not necessarily be the most relevant or representative.

When we talk about accessibility bias in language models, it can stem from training on large amounts of publicly available data. As a result, the model is more likely to favor content that is more accessible (e.g., popular culture and current events). In contrast, it neglects perspectives and information that are less prevalent online. For example, training data is likely based on freely available text, not content hidden behind paywalls. It can impact access to some perspectives in the training sets.

Availability bias can also manifest in how the language model generates responses. If certain information or expressions are more prominent in the training data, the model may create responses consistent with that information. It's not necessarily a direct bias in the data itself but rather an inequality in how the model generates responses based on the most "accessible" or prominent patterns.

Accessibility bias in a language model can create information bubbles and echo chambers that reinforce existing biases instead of fostering different perspectives. It can also lead to misinformation about a given topic if that misinformation is easier to get hold of than factual content.

Representation bias

Representational bias in a language model refers to the situation where training data does not correctly represent the domain for which the model is trained. Language models, such as ChatGPT, are trained on a vast data set, which, in an ideal situation, reflects the entire world as a domain and, therefore, has an objectivity and diversity that ensures that all aspects of a topic are elucidated. For example, if topics are not represented in training data or the subject is not illuminated from all angles, it can cause representation bias. If all texts in training data on a particular topic are written by one demographic group, there may be a lack of perspectives and opinions from other groups. When data comes solely from publicly available Internet data, such as the Internet, it is impossible to access. This selection will also have a representation bias in Wikipedia, news articles, internet forums, and selected books. We usually place a higher value on peer-reviewed scientific articles than free content from the Internet. Such scientific articles are often behind paywalls and are probably not included in training datasets. If training data is old and the world has changed since training the model, the model may have a representation bias. (For example, ChatGPT doesn't know anything about the war in Ukraine. It only knew about Russia's invasion of Crimea when it finished training in September 2021).

Historical bias

Historical bias can also have an impact on language models. Historical texts in training data can reflect norms, values, and attitudes from when they were written. These may differ from today's norms and may be skewed or discriminatory. There may also be historical situations where specific demographics have been marginalized or omitted from historical records. Moreover, historical sources may contain factual errors or distorted interpretations of events, so these perspectives are reflected in the language model.

Selection bias)

Selection bias occurs when the training data is not representative of the entire population or target audience. If specific perspectives are underrepresented or excluded from the training data, the AI model will lack the necessary "knowledge" to generate unbiased and utterly objective content.

For example, if the training data primarily includes data from Western countries, the language model cannot produce accurate and culturally relevant content for non-Western audiences. This omission perpetuates societal inequalities and prevents the model from being an inclusive and objective source of

information. Another challenge may be that training data comes from specific sources (e.g., news websites, scientific articles, or social media). This can lead to a skewed understanding based on the perspectives and styles typical of these sources. Most language models that are freely available, such as S.C., are not available languages. ChatGPT is primarily trained on Western (American) texts, which is why it will incline Western (American) culture and Western (American) values. China has, therefore, created its language model and chatbot, called Ernie Bot, which is made so that it is true to the Chinese regime and its values.

Group attribution bias

Group attribution bias occurs when the generative AI (applicable to both language models and imaging models) attributes specific characteristics or behaviors to an entire group based on the actions of a few individuals. For example, the models may associate negative attributes with specific ethnicities or genders, perpetuating harmful generalizations and prejudices. To avoid this, the models must be trained on different data sets that reflect the complexity and individuality of other groups.

Contextual bias

Contextual bias is defined as biases that occur when one relies on one particular context to get a correct answer. It happens when language models have challenges in "understanding" or "interpreting" the context of a conversation or prompt accurately. Misunderstanding the context can lead to the generation of inappropriate or misleading responses. Although the significant language models can generate responses, they lack a deep understanding of context. This can lead to technically correct answers but lack nuance or relevance in the given situation.

ChatGPT has a minimal "memory" in terms of past interactions in a conversation. It often cannot "remember" previous conversations (chat sessions). Suppose a user provides important context early in a conversation. In that case, the model may not always remember or consider that context in later replies so that the user may receive an unexpected response. Unlike humans, who can ask clarifying questions if they are unsure of the context, language models generate answers based solely on the information it has available. It can lead to misunderstandings or inaccurate answers if the context is not clearly stated in the prompt. If the prompt lacks specific context about the available data, the language model can produce generalized or stereotypical answers. ChatGPT is built to almost always provide an answer – whether the answer is correct or not.

A possible solution to this problem lies as much in the fine-tuning of the language model by human feedback as in the unsupervised training on data itself. When ChatGPT allows the user to provide input on whether a response is appropriate for a given prompt by thumbs up/thumbs down, it is, among other things, to minimize contextual bias.

Linguistic bias

Linguistic bias deals with the biases inherent in language itself or in the way language is used. These biases can reflect and reinforce cultural, social, or cognitive biases. If the texts in training data for language models contain linguistic trends, the model is likely to do the same. For example, if certain adjectives are often used

in conjunction with specific nouns (e.g., "strong man" vs. "caring woman"), the model can recreate these associations in its outputs. If the training data contains stereotypical or prejudiced expressions or phrases, the language models will also use or confirm these expressions in their responses. Some languages have grammatical structures that may reflect cultural biases (e.g., gender inflection of adjectives or nouns). If these structures are skewed in training data, the model can reproduce these biases.

If training data is dominated by content in a particular language (e.g., English), the model may be better and more nuanced than other languages. This can lead to an imbalance between the dominant language and the content in different languages being deprioritized. In addition, language models will often use machine translation into English and back. Here, too, imbalances can arise.

Anchoring bias)

Anchoring bias refers to people's tendency to rely too much on the first piece of information they receive when making decisions. After receiving an "anchoring," people adjust their subsequent judgments and decisions based on that anchoring, even if it may be irrelevant or erroneous. When a user asks a question to a language model, it will use that input as an anchor and generate an answer based on it. Suppose the user's prompt contains a specific bias or angle. In that case, the model may create responses that follow that bias, even if it is not necessarily the most objective or accurate response.

People can reconsider and adjust their initial and immediate assessments after receiving more information. Language model responses are based on patterns in training data and the immediate context of the user's prompt. This means the language model does not actively "reconsider" previous information in a chat session but instead responds to the latest anchor (prompt) it receives.

It's important to note that while humans may fall for anchoring bias due to cognitive biases, this kind of bias in language models occurs primarily because of its interface design and its training data. The model responds to the immediate input and generates responses based on this alone.

Automation bias

Automation bias refers to the tendency of humans to rely excessively on automated systems and machines, often at the expense of human judgment and common sense. When humans rely too much on machines, they may overlook errors or inaccuracies produced by the automated system and fail to intervene when necessary.

Users may rely too much on answers generated by language models like ChatGPT because they assume the answers are always correct. After all, language models seem enormously convincing and authoritative when they write. Many would also consider ChatGPT to be an actual authoritative source, although it often makes mistakes and gives inaccurate answers. Due to the impressive ability of modern language models to generate persuasive and often correct-sounding responses, many users will fail to critically evaluate the answers or seek further confirmation (source criticism). When a prompt isn't precise enough, or the language

model doesn't "know" what is being asked, the system guesses as qualified as it can. These guesses are based on statistics and probability and will, therefore, lead to biases - and thus often completely wrong answers!

At the same time, many users will become too dependent on artificial intelligence's ability to generate content, answer questions, or make decisions, which can lead to an underestimation of human expertise or intuition. In this way, our ability to be critical of our surroundings can be affected. Users must know about automation bias when interacting with ChatGPT and other automated systems. Although ChatGPT is an advanced language model, it is far from flawless and should not replace human judgment. Users should constantly critically assess the model's response and seek further confirmation when necessary.

Confirmation bias

Affirmation bias is a psychological mechanism in which individuals seek and remember information that confirms their existing beliefs while ignoring knowledge that challenges them.

CONFIRMATION BIAS



Source: <https://www.simplypsychology.org/confirmation-bias.html>

Suppose training data on which language models are based contain confirmation bias from human sources (e.g., skewed or tendentious news articles, blogs, or forums). In that case, this bias may become embedded in the model. This means that if a particular belief or viewpoint is overrepresented in training data, the model may tend to generate responses that confirm that view.

Language models do not critically evaluate information. It generates responses based on statistics and patterns in training data. If a user asks a question with a specific angle, the model may create an answer confirming that angle because it reflects the patterns seen in training data. Similarly, language models may tend to generate responses based on what is best represented in training data rather than what is necessarily accurate or objective. This can lead to the model confirming popular, but perhaps erroneous, beliefs.

It's important to understand that while humans can fall for confirmation bias due to cognitive biases, this kind of bias in language models occurs primarily because of its training data and the way it's designed to generate responses. Users should be aware of this limitation and take the model's reactions with a grain of salt, especially on controversial or sensitive topics.

Fine-tuning language models also add bias!

Language models like ChatGPT undergo fine-tuning by a process called reinforcement learning with human feedback (RLHF). In this process, people who assess the model's responses are used and then adjust the model to respond better to human norms and values. The big challenge is that people have very different norms and values based on their origins, among other things. ChatGPT will likely respond regarding American ethics, norms, and values as an American company trains the model in the United States.

There are certainly more possible types of bias in language models than the ones I've described in this article. The idea is not to make a 100% inexhaustible list of potential biases that may arise when working with generative artificial intelligence but to focus on and inform about the challenges.



I am not taking a position on *whether* there are biases or, if so, what kinds of biases are present in the different language models, but have only described a number of potential challenges.

Biases in a Google search

Most people will probably see a Google search as an excellent alternative to using ChatGPT to find information and write about a topic. But is it better and safer to use? Google personalizes search results based on the user's past search history, web history, location data, and other personal data. This means that two different users can get extra search results for the exact search. While this can improve the relevance of the results for the individual user, it can also create a "filter bubble" where the user only sees information that confirms their existing beliefs. Google is a commercial company, in the same way, that OpenAI, which is behind ChatGPT, is, and some of their algorithms will prioritize paid ads and other commercial interests over organic content. The top (and thus anchoring) results are most often ads that some have paid to get favored! Algorithms can also prioritize content expected to increase user engagement, highlighting more sensational or controversial content over more nuanced or objective content. Google applies different quality criteria to assess the credibility and authority of other websites. One might think this is a good thing, but in some cases, it can lead to lesser-known or alternative views being downgraded.

A Google search emanating from the user may also cause bias. Users tend to search for topics and click on links confirming their beliefs, i.e., confirmation bias. This, too, can affect the personalized results that the user will see in the future. The way a question or search is formulated can affect the results you get. For example, a search for "the advantages of X" can yield very different results than a search for "the disadvantages of X". Many users check only the first few results on a search page. This can mean that they miss out on more in-depth or alternative perspectives that might be further down the page, and most often, they'll only see the paid search results! Users also tend to rely on sources they already know or consider authoritative and may overlook or underestimate information from lesser-known sources.

The main difference between a Google search and a ChatGPT writing about a topic is that you often write and compose your content from multiple sources when you search on Google (human in the loop). In contrast, you let ChatGPT write itself, without knowing where the information comes from.

Humans also have biases!

All people are biased in a multitude of ways. When we meet a new person, we have a lot of prejudices (conscious or unconscious) on which we judge the person. We consciously or unconsciously notice superficial aspects such as gender, ethnicity, age, appearance, dress, body adornments, etc. We use this information intuitively to judge the person. We often have some expectations in advance that dominate our impression of the person. We also seek comparisons between, for example, people we know and new people we meet. We may tend to attribute characteristics to new people in our social circle based on our ideas based on comparisons. We also tend to like people who look like ourselves or who we know better than some who are different.

Human beings will tend to confirm bias, that is, seek, interpret, and remember information in a way that confirms one's existing beliefs or hypotheses. We have an anchoring bias. We rely primarily on the first information we get. We succumb to availability bias, where we base our judgments on information that is readily available or fresh in our memory rather than relevant information. We attach more credibility and allow ourselves to be influenced more by authorities or experts, and we tend to do the same as the majority do. Many people prefer to do as usual rather than change or innovate, and most favor their abilities over those of others. And then there's the Dunning-Kruger effect: the less you know about a subject, the more confident you are in your case!

Authority Bias

We are more likely to trust and be influenced by ideas that come from authority figures



"Our CEO says the state might tax disposable cups". "Hey, I told you that last week and you said it would never happen!"

Confirmation Bias

This occurs when we warp data to fit or support our existing beliefs or expectations



"What the human being is best at doing is interpreting all new information so that their prior conclusions remain intact."

Sunk Cost Bias

We are often influenced by past, sunk costs, which continue to distort our decisions



"I have already spent so much time and money on this project, so I might as well keep going"

Halo Effect

Our overall impression of a person influences how we feel and think about his or her character



"I think she's too nice to be a good commander in chief"

Availability Cascade

An idea accumulates more credibility as it spreads



"I've heard from a bunch of people that he's going to raise taxes, there must be some truth to it"



Dunning-Kruger Effect

The less you know, the more confident you are



"Our CEO says the state might tax disposable cups". "Hey, I told you that last week and you said it would never happen!"

Declinism

We romanticize the past and believe that society and institutions are in decline



"The latest Supreme Court decision is just another example of how our country is falling apart these days"

Framing Effect

We draw different conclusions based on how an idea is presented to us



Doctor A: "You have an 80 percent chance of a full recovery."
Doctor B: "There's a 20% chance that you'll die after being treated"

Bandwagon Effect

Conforming to a widely held world view in order to fit in and minimize conflict



"All my friends and colleagues are buying that new cryptocurrency. The price keeps on rising, I must buy it"

False Consensus

Overestimating the proportion of people who agree with an idea



"I think that the majority of people agree that this policy makes sense. Everyone I know thinks so"

Source: <https://mahanakornpartners.com/the-effect-of-cognitive-bias-in-decision-making/>

We also have lots of gendered biases in our language. The Danish Language Council has made a census, and in the Danish language, there are 187 words ending in '-man,' while only 14 words end in '-woman.' Most are job titles, but that tells a lot about linguistic bias. We have previously written about "gender bias" and

artificial intelligence in the article "Gender bias when generative AI writes texts". The Danish Language Council will clean up the gendered designations when the new edition of the spelling dictionary is published in 2024.

All these biases and many more will be reflected in the texts we write, and therefore, they will be transferred to the language models trained on artificial texts.

We can also discuss whether we, as teachers, are 100% free of bias when selecting materials for students and when we teach. Are all educators utterly free of prejudices and preferences? Perhaps it is precisely when differences meet that good learning occurs – at least as long as we are aware of our own biases. And what does this mean about the bias of language models?

What can we do about biases in language models?

The most important thing we can do is be aware of possible biases, whether human bias, search engine bias, or artificial intelligence. We must recognize that they can influence our decisions and actions in many different ways. We must be critical of the information we receive - conscious as well as unconscious. This is especially true for information from language models such as ChatGPT, as the models lack transparency. We have no idea where the data comes from or if it is correct. We don't know how the datasets are composed. We don't understand why the language model chooses the words it does. We have no sources to check, and we are easily overwhelmed by the convincing and very credible texts that the machines write.

As educators, we must enlighten and educate our young people. It is essential that we make future generations critically thinking and reflective people and that they learn to use the new technologies on an informed basis. **Informing and teaching about the potential biases of technologies, the need for critical thinking, validation of information, and healthy skepticism, might be a good place to start!**

In any case, we must be very conscious of what we can *expect* from language models when we use them!



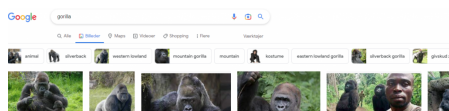
Biases are one of the main focuses of the developers of significant language models! It's just tough to get rid of, as it's everywhere. If we were to train a language model without bias, we first had to define what the ideal world is, seen through the eyes of all people. You would have to make a "synthetic dataset" that was validated and 100% bias-free. Training data for models like ChatGPT is many billions of sentences, so this task is virtually impossible. And since people are different, there will always be an imbalance in some people's eyes when others see a perfectly balanced model. We humans are influenced by the culture and society we live in, and therefore, the only solution is perhaps to create completely *local language models* that fit just that.

Sources:

Det er menneskeligt at fejle – derfor gør teknologier det også

KOMMENTAR: Algoritmer tilegner sig ofte fejl, baseret på skævheder i de input, de trænes ud fra. Løsningen er både menneskelig og teknisk.

 Videnskab.dk Christina Lioma



Is ChatGPT Woke And Biased? 16 Examples That Prove So

Is ChatGPT Woke And Biased? We've curated the latest as well some popular examples from the internet that reveals Chat GPT Wokeness And Biasness.

 Insane - TheInsaneApp.com Editorial Staff



https://medium.com/@savusavneet_28467/biases-in-machine-learning-ml-32b147492242

Council Post: Navigating The Biases In LLM Generative AI: A Guide To Responsible Implementation

Prudent utilization of LLM generative AI demands an understanding of potential biases.

 Forbes Ken Knapton



<https://rockcontent.com/blog/chatgpt-bias/>

<https://openwebtext2.readthedocs.io/en/latest/background/>

<https://arxiv.org/pdf/2005.14165.pdf>

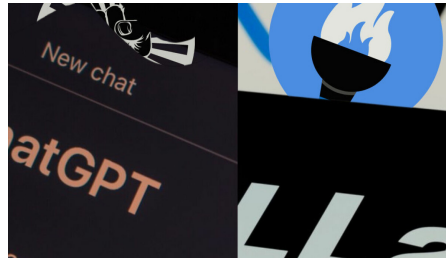
<https://www.sciencedirect.com/science/article/pii/S266734522300024X>

<https://rockcontent.com/blog/chatgpt-bias/>

Er ChatGPT venstreorientert? Forskere har undersøkt de politiske holdningene til 14 chatboter

ChatGPT var den mest venstreorienterte, mens Metas chatbot lå lengst til høyre: Ny studie kaster lys over chatbotenes politiske ståsted.

 [Forskning.no](#) Frederik Guy Hoff Sonne




Algorithm Bias In Artificial Intelligence Needs To Be Discussed (And Addressed)

You have a part to play in the matter...

 [Towards Data Science](#) Richmond Alake

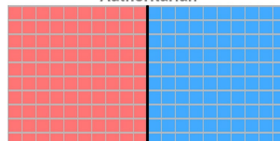
The political orientation of the ChatGPT AI system

Applying the Pew Research Political Typology Quiz to a state-of-the-art AI Language model

 [Rozado's Visual Analytics](#) David Rozado

Results of applying the Political
Compass Test to ChatGPT

Authoritarian



How to remove bias from your AI-generated content - DEPT®

AI tools can create biased output. Learn how to recognise and prevent it, in order to easily create content that reflects modern society.

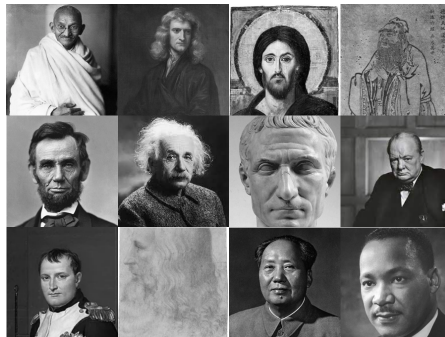
DEPT®Michelle den Elzen Team Lead Copywriting



Where Are All the Women?

Exploring large language models' biases in historical knowledge

Towards Data ScienceYennie Jun



The Hidden Biases in ChatGPT: What You Need to Know

Investigate the inherent biases in machine learning models like ChatGPT and how they can impact the information you receive.

MediumJohnny | AI Bites



The politics of AI: ChatGPT and political bias | Brookings

When asked to indicate support or lack of support for a variety of political statements, ChatGPT's responses tend to replicate a liberal point of view, albeit with logical inconsistencies, emblemizing the issue of bias embedded in AI systems through their datasets and human trainers.

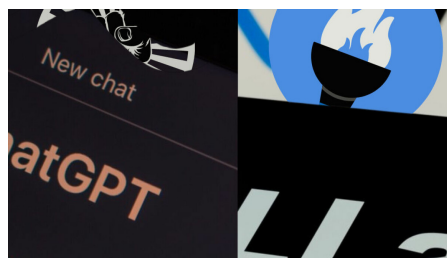
 Brookings_jeremybaum



Er ChatGPT venstreorientert? Forskere har undersøkt de politiske holdningene til 14 chatboter

ChatGPT var den mest venstreorienterte, mens Metas chatbot lå lengst til høyre: Ny studie kaster lys over chatbotenes politiske ståsted.

 Forskning.noFrederik Guy Hoff Sonne



Snydt af din egen hjerne: Du tror mest på fakta, der passer til dine holdninger

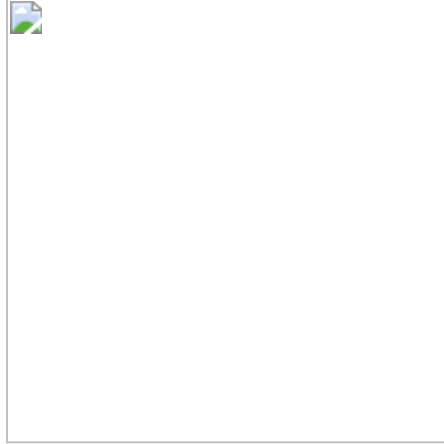
Videnskab.dk's hjernepodcast 'Brainstorm' dykker ned i fænomenet kognitiv bias.

 Videnskab.dkFrederik Guy Hoff Sonne

Videnskabsteori - Bias

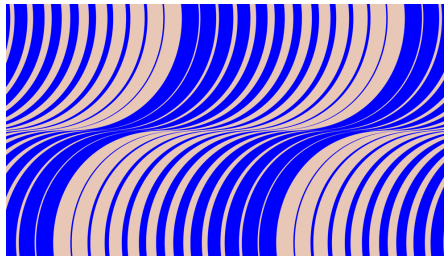
Systematiske fordrejninger i vores tænkning

 Bias




How should AI systems behave, and who should decide?

We're clarifying how ChatGPT's behavior is shaped and our plans for improving that behavior, allowing more user customization, and getting more public input into our decision-making in these areas.



Exploring the potential for bias in ChatGPT - National centre for AI

As part of our work at Jisc we want to help institutions adopt AI in an ethical and responsible way, and understanding bias is an important part of this. Here we explore the potential for bias in ChatGPT.

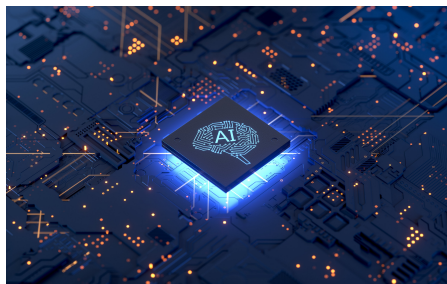
 National centre for AI Michael Webb



Historical bias in AI systems

Learn more about how historical bias in data sets can negatively impact developing and deploying ethical AI.

 Australian human rights commission



How OpenAI is trying to make ChatGPT safer and less biased

Plus: AI is dreaming up drugs that no one has ever seen. Now we've got to see if they work.

 MIT Technology ReviewMelissa Heikkilä



Formand eller forkvinde? Mændene dominerer vores sprog

Overenskomstforhandlinger er i gang i hele landet, og her ser man ordet 'forligskvinde' blive brugt - i stedet for 'forligsmand', som historisk er mere brugt.


 DRMona Aaberg

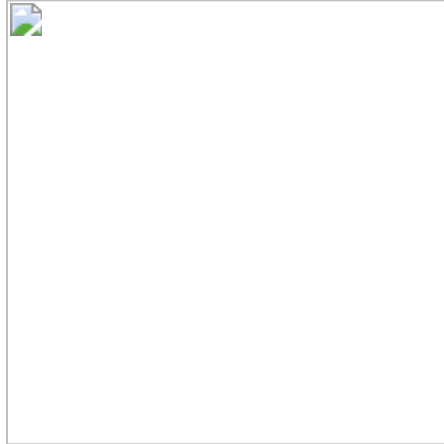
forman 

<https://doi.org/10.48550/arXiv.2304.03738>

Towards Understanding and Mitigating Social Biases in Language Models

As machine learning methods are deployed in real-world settings such as healthcare, legal systems, and social science, it is crucial to recognize how they shape social biases and stereotypes in the...

 PMLR Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, Ruslan Salakhutdinov



Redirecting

E