# Biases and stereotypes in generative AI: A discussion guide

Generative artificial intelligence has much potential in teaching and learning. It also has a dark side that students and faculty need to understand and explore. This includes biases in language, [data](https://cmr.berkeley.edu/2020/11/algorithmic-bias/), and [algorithms](https://en.wikipedia.org/wiki/Algorithmic_bias); a tendency to fabricate; a potential to take on a threatening persona, and an origin story with disturbing elements. As with other technologies, these biases are amplified when an automated system replaces human judgment, as in such things as screening in job applications and college admission, and security screenings and searches.

Bias is an inherent part of who we are, what we know, and what we have experienced. There is no way to eliminate it. By better understanding the many facets of bias and considering how it influences our actions, we can improve our ability to lessen its effects. That’s what this handout is for. It is the first of a series of guides intended to help instructors and students analyze and discuss the biases and ethics of generative AI, or, more specifically, large language models like ChatGPT, Bing Chat, Dall-E, and Midjourney that produce written work, images, and code from natural language prompts.

## Biases in AI

ChatGPT, Bing Chat, and other AI chatbots are trained on an enormous amount of publicly available online text. As a result, they have the same biases as those sources and the society that produced them. These include:

* They are dominated by a white, male perspective.
* They are highly influenced by American culture, American capitalism, and the English language.
* They can and will invent sources, people, and events. (The industry term is “hallucinate.”)
* They can generate offensive responses, although filtering has reduced the likelihood that will happen.
* AI detectors are more likely to flag the work of writers for whom English is not a first language.

None of that means we should avoid AI tools. Rather, we need to remember that human judgment is crucial. Chatbots cannot reason or make decisions. The writing they produce is based on predictions of word order gleaned from the sources they have analyzed. As with any other information, we need to evaluate AI-created material with a critical eye.

## Behind the biases

Programmers are well aware of the biases that emerge in generative AI. [As Emilio Ferrara writes](https://www.llrx.com/2023/07/eliminating-bias-in-ai-may-be-impossible-a-computer-scientist-explains-how-to-tame-it-instead/), though, it is difficult to determine where bias in large language models originates: The data used for training? The algorithms used for training? Human reviewers who provide feedback during testing? Audiences that programmers prioritize (corporate over general, for instance)? Or did the “guardrails” that companies created to cut down on bias and offensiveness lead to different types of biases? [Any or all of those things could be at fault](https://arxiv.org/abs/2304.03738).

Addressing bias in generative AI systems is both challenging and problematic, Ferrara says. Training data and model configurations can be changed, but the results are impossible to predict. Additional programming that directs the AI model to avoid certain words or phrases could lead to misleading responses or make systems less useful.

Language itself contains biases, Ferrara says, and it is difficult to separate appropriate use in one cultural context from inappropriate use in another. Language patterns “are often deeply ingrained in the way people express themselves and the structures of language itself,” Ferrara writes. ChatGPT and other large language models are trained on text from websites, books, social media, and online chats, so biases in that material are inevitably reflected in outputs from chatbots.

Another challenge in reducing biases in generative AI is determining the values AI should follow. Jackie Davalos and Nate Lanxon of Bloomberg News [explain it this way](https://www.bloomberg.com/news/videos/2023-06-29/the-alignment-problem-ai-irl-episode-11): “We’re building systems and we’re saying they’re aligned to whose values? What values?” The author Brian Christian calls this [an “alignment problem.”](https://blogs.nvidia.com/blog/2021/02/03/brian-christian-alignment-problem/)

Ferrara argues that instead of trying to eliminate biases in generative AI, we should work toward fairness and an alignment with human values. Of course, the terms *fairness* and *bias* are ambiguous in their own right, given the wide range of perspectives and beliefs in societies. Ferrara lists several types of bias:

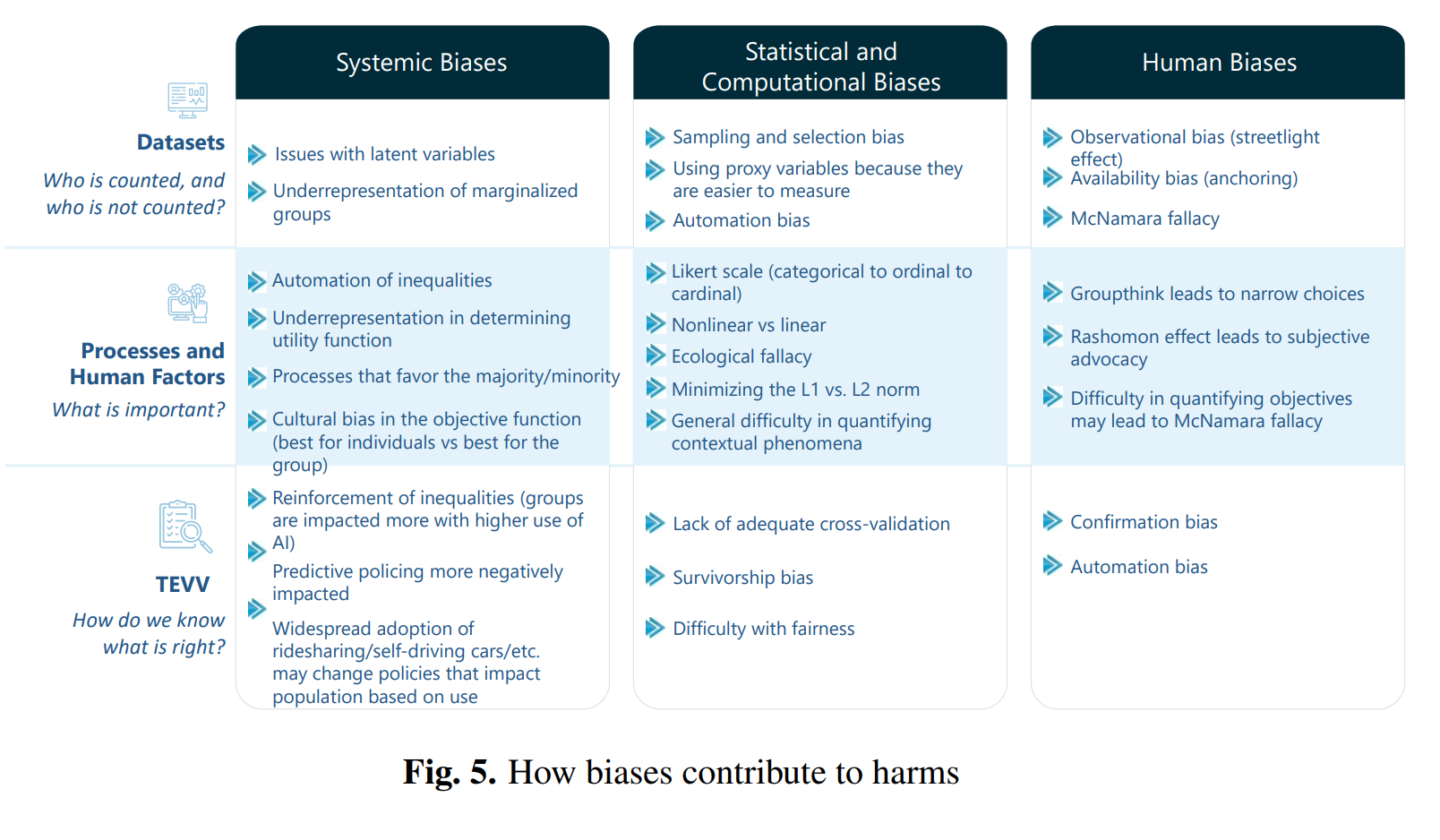
* **Demographic biases**. Training of language models may over-represent or under-represent characteristics of race, gender, ethnicity, and social groups.
* **Cultural biases**, including prejudices and stereotypes.
* **Linguistic biases**. English and a few other languages are most prevalent in online content, so generative AI models mostly reflect those languages.
* **Temporal biases**. Generative AI models are trained on material from particular time periods, so they may lack material before or after a certain point in time.
* **Confirmation biases**. Generative AI may reinforce flawed ideas or biases by providing only the views or representations a user expects.
* **Ideological and political biases**. The responses from generative AI will reflect political and ideological biases from training materials.

**Other types of biases**

A [resource from the National Institute of Standards and Technology](https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf) provides additional perspectives on bias. These include:

* **Statistical**: All statistical models and technological systems require assumptions from researchers and system designers. “These decisions affect who and what gets counted, and who and what does not get counted,” the report says.
* **Dataset**.It is easiest to use available data than to seek out data that better represents a general population.
* **Cognitive and societal**. These include biases in decisions about which data to use in developing AI, how AI will be used, who should pay for it, and how it will be made available – and to whom.
* **Systemic**. Long-used procedures and practices that favor some people and disadvantage others. “This need not be the result of any conscious prejudice or discrimination but rather of the majority following existing rules and norms.”
* **Human**. These are inherent in all other biases, including perceptual bias, anchoring bias, confirmation bias, and framing effects.
* **Techno**-**solutionism**. A belief that technology is crucial to solving many of the world’s problems.

The charts below and another on the last page provide a good synopsis of the types of biases to consider. Both are from the National Institute of Standards and Technology report. The report’s glossary includes other biases to consider.



From [Towards a Standard for Identifying and Managing Bias in Artificial Intelligence](https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf) (2022).

**Biases of individualism and collectivism**

Two other social and cultural aspects to consider with bias are [individualism and collectivism](https://eml.berkeley.edu/~groland/pubs/IEA%20papervf.pdf).

* **Individualism**. This is the foundation of American culture, capitalism, and academia, with an emphasis on personal freedom, personal choice, recognition of individual achievement, and the accumulation of wealth.
* **Collectivism**. This is common in East Asia and other societies, and values conformity and the well-being of society over the individual.

Because we live in an individualistic culture, we often overlook the way this view shapes how we approach many aspects of life, including technology use. For instance, many Americans object to companies training generative AI on online digital materials, something that isn’t an issue in collectivist societies.

## Class discussion

Here are a few ideas for discussions and activities in which students explore the biases and other problems in generative AI.

### Biases in language

1. Create a prompt that all students must use to generate text with a chatbot, but have them try different bots: ChatGPT, Bing Chat, Bard, Claude 2.

* What similarities and differences do they see in what the bots produced? What strengths and weaknesses?
* What types of biases are reflected?
* How would they rate the quality of the writing?
* Did the bots provide sources? If so, are they real? Accurate?

1. Give students a variety of prompts or have them create their own prompts related to class material. Alternatively, you can use generative AI to create material for students to evaluate.

* What biases or stereotypes, if any, can students identify in the bot-produced work?
* How could students change their prompts to avoid the types of bias they see?

1. [One group of researchers found](https://blog.allenai.org/toxicity-in-chatgpt-ccdcf9265ae4) that having ChatGPT assume a personality increases the chances of toxic responses substantially. That toxicity varies widely depending on the persona. For instance, a persona of a journalist is likely to produce more toxic results than a persona of a businessperson. The persona of Richard Nixon is more toxic than that of John F. Kennedy. The researchers warn that “ChatGPT can be nudged to generate harmful content, especially when assigned a persona.”

* Have students assign different personalities to ChatGPT and other chatbots and then ask similar questions.
  + How do the responses differ?
  + Are there noticeable differences in tone or toxicity?
  + What might cause these types of responses from generative AI?

1. Lynne Walsh, a journalist and assistant director of the nonprofit site Trusting News, [suggests using generative AI](https://mailchi.mp/afd673511e1b/trust-tips-224-can-chatgpt-and-ai-help-us-prevent-bias-and-polarization-in-our-reporting) to spot biases in unpublished work. She focuses on journalistic writing, but the same approach could be used in academic writing or any other type of writing.

* Have students use a generative chatbot to evaluate an article for biases. It could be their own work or an article they find online.
  + Possible prompt: *Read this article and identify potential problems in (*identify the types of biases you want to focus on*). Create a list of those problems, explaining what each problem is and how it might be resolved.*
* What sorts of biases, if any, do the chatbots flag?
* Do students see value in trying this with their own work in the future?



### Visual biases

The biases, assumptions, and stereotypes of generative AI are easy to see in AI-produced images. In the examples above, professors are depicted as white and male. A prompt for nurses as superheroes returned images of buxom women (mostly white) in short skirts and tight tops. Those are just two examples. Here are some other ways of helping students explore biases of AI through images.

**Exercises and class discussions**

1. Create a prompt that all students must use in creating images with AI image creators: [Dall-E](https://openai.com/dall-e-2), [Stable Diffusion](https://stablediffusionweb.com/), [Bing Image Creator](https://www.bing.com/create), [Night Café](https://creator.nightcafe.studio/), [Catbird](https://www.catbird.ai/) (all are free but require creation of an account).

* What types of biases or stereotypes do they see in the images?
* How can you alter the prompts to reduce or eliminate the stereotypes or biases?

1. Have students use different AI models to create images of different types of people or professions. For example: students, nurses, professors, cooks, chefs.

* What types of biases do you see in the representations? In their surroundings?

1. Have students create images of different races, ethnicities, and genders.

* How are people represented?
* How well are the images rendered? (Many AI image models have difficulties with non-white faces.)
* How could they change prompts to create more accurate representations?

1. Have students view these [AI-generated images of people from each U.S. state](https://abc3340.com/news/offbeat/gallery-ai-shows-a-couple-from-each-state-if-they-were-dressed-in-that-states-style-artificial-intelligence-microsoft-image-creator-dall-e-california-florida-texas-new-york-pennsylvania-oregon-utah-tnd-photos).

* What types of stereotypes or biases, if any, do they see in the images?
* Where do these stereotypes originate?
* If these are stereotypes, how would students render the same types of images?
* Compare these to other representations created by the same site:
  + [Couples from the year 3000](https://thenationaldesk.com/news/from-the-desk/gallery-ai-imagines-what-each-states-average-couple-year-3000-artificial-intelligence-future-futurism-futuristic-space-people-aliens-photo-gallery-typical-stereotypical-couple-man-woman-california-florida-new-york-texas-utah-oregon).
  + [Couples from each state if drawn by Pixar](https://thenationaldesk.com/news/from-the-desk/gallery/gallery-ai-shows-us-if-the-average-couple-in-each-state-was-drawn-by-pixar-disney-finding-nemo-dory-monsters-inc-toy-story-cars-incredibles-brave-elemental-inside-out-coco-moana-animation-animators-artificial-intelligence-photo-galleries-tnd?photo=1).

1. Have students read the article below.

* [AI and the American Smile: How AI misrepresents culture through a facial expression](https://medium.com/@socialcreature/ai-and-the-american-smile-76d23a0fbfaf), by Jenka. Medium (26 March 2023).
* Do they agree with the author’s premise that the images in the article represent an “American smile”? Why or why not?
* Have students use generate AI to create similar types of group images. Do they see similar representations of what the author calls “a lying smile”?
* What other cultural distortions have they seen in AI-generated images?

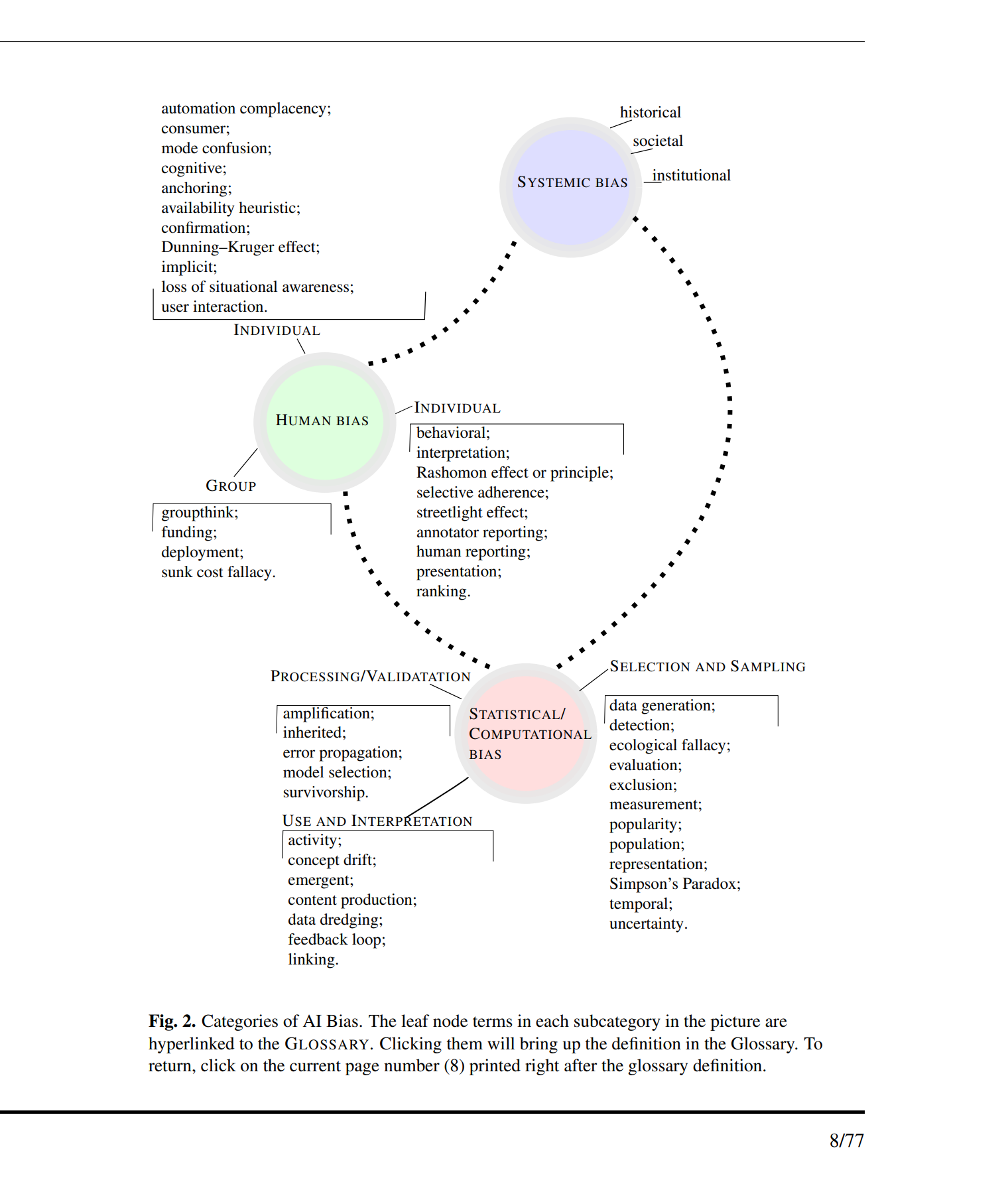
## Readings and resources

This is a highly selective list of resources you might draw on in helping students discuss the biases and stereotypes they are likely to encounter in using generative artificial intelligence.

* [A Blueprint for an AI Bill of Rights for Education](https://criticalai.org/2023/07/17/a-blueprint-for-an-ai-bill-of-rights-for-education-kathryn-conrad/), by Kathryn Conrad. Critical AI (17 July 2023).
* [AI language models are rife with different political bias](https://www.technologyreview.com/2023/08/07/1077324/ai-language-models-are-rife-with-political-biases/), by Melissa Heikkila. MIT Technology Review (7 August 2023).
* [Artificial Intelligence: Trust Issues](https://www.linkedin.com/pulse/artificial-intelligence-trust-issues-kerry-johnson/), by Kerry Johnson. LinkedIn (13 April 2023).
* [AI writing assistants can cause biased thinking in their users](https://arstechnica.com/science/2023/05/ai-writing-assistants-can-cause-biased-thinking-in-their-users/), by Elizabeth Rayne. Ars Technica (26 May 2023).
* [Can ChatGPT and AI help us prevent bias and polarization in our reporting](https://mailchi.mp/afd673511e1b/trust-tips-224-can-chatgpt-and-ai-help-us-prevent-bias-and-polarization-in-our-reporting), by Lynn Walsh. Trusting News newsletter (27 June 2023).
* [ChatGPT’s Fluent BS Is Compelling Because Everything Is Fluent BS](https://www.wired.com/story/chatgpt-fluent-bs/), by Amit Katwala. Wired (9 December 2022). *“ChatGPT was trained on real-world text, and the real world essentially runs on fluent bullshit,” from media to business to politics to education.*
* Concepts of Ethics and Their Application to AI, by Bernd Carsten Stahl. (2021). In [Artificial Intelligence for a Better Future](https://link.springer.com/book/10.1007/978-3-030-69978-9). SpringerBriefs in Research and Innovation Governance. Springer, Cham.
* [The Ethics of ChatGPT: Ensuring AI Responsibly Serves Humanity](https://readwrite.com/the-ethics-of-chatgpt-ensuring-ai-responsibly-serves-humanity/), by Micheal Chukwube. ReadWrite (24 June 2023).
* [Even artificial intelligence can acquire biases against race and gender](https://www.science.org/content/article/even-artificial-intelligence-can-acquire-biases-against-race-and-gender), by Matthew Hutson. Science (13 April 2017).
* [Fresh Evidence of ChatGPT’s Political Bias Revealed by Comprehensive New Study](https://www.uea.ac.uk/news/-/article/fresh-evidence-of-chatgpts-political-bias-revealed-by-comprehensive-new-study). University of East Anglia (17 August 2023).
* [Gender Bias in Futuristic Technologies: A Probe Into AI & Inclusive Solutions](https://www.orfonline.org/research/gender-bias-in-futuristic-technologies-a-probe-into-ai-inclusive-solutions/), by Mona. Observer Research Foundation (17 January 2022).
* [GPT detectors are biased against non-native English writers](https://doi.org/10.48550/arXiv.2304.02819), by Weixin Liang, et al. *Computation and Language* (6 April 2023).
* [Humans Are Biased. Generative AI Is Even Worse](https://www.bloomberg.com/graphics/2023-generative-ai-bias/), by Leonardo Nicoletti and Dina Bass. Bloomberg, (n.d.). An excellent analysis of AI-generated images of people. The interactive graphics make the data easy to comprehend. If you run into a paywall, [try this version of the article without the graphics](https://archive.is/2WEsN)
* [Rise of the racist robots – how AI is learning our worst impulses](https://www.theguardian.com/inequality/2017/aug/08/rise-of-the-racist-robots-how-ai-is-learning-all-our-worst-impulses), by Stephen Buranyi. The Guardian (8 August 2017).
* [Teaching AI Ethics](https://leonfurze.com/2023/01/26/teaching-ai-ethics/), by Leon Furze. Blog post (26 January 2023). A good resource with many questions and sources. It includes a graphic about the ethical considerations: bias, environment, truth, copyright, privacy, datafication, affect recognition, human labor, power. (Also see a PDF handout on the site about ethics.)
* [These robots were trained on AI. They became racist and sexist](https://www.washingtonpost.com/technology/2022/07/16/racist-robots-ai/), by Pranshu Verma. Washington Post (16 July 2022).
* [Towards a Standard for Identifying and Managing Bias in Artificial Intelligence](https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf), by Reva Schwartz, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt, and Patrick Hall. NIST Special Publication 1270. National Institute of Standards and Technology (March 2022).
* [Two Truths and a Lie](https://thebrainyacts.beehiiv.com/p/114-two-truths-lie), by Josh Kubicki. The Brainacts 114 (2 August 2023). *The relevant information is in the first part of the post.*
* [Why algorithms can be racist and sexist](https://www.vox.com/recode/2020/2/18/21121286/algorithms-bias-discrimination-facial-recognition-transparency), by Rebecca Heilweil. Vox (18 February 2020).

**Resources**

* [Diffusion Bias Explorer](https://huggingface.co/spaces/society-ethics/DiffusionBiasExplorer). A free interactive tool for looking at how the visual-creation models Stable Diffusion and DALL-E represent more than 100 professions. You can also combine the professions with about 20 different adjectives.
* [Gender Shades](http://gendershades.org/). Joy Buolamwini of the MIT Media Lab talks about the gender and racial biases in facial recognition systems. You can also read a [related research article](http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf) and a [multimedia document](http://gendershades.org/overview.html).



From [Towards a Standard for Identifying and Managing Bias in Artificial Intelligence](https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf) (2022).