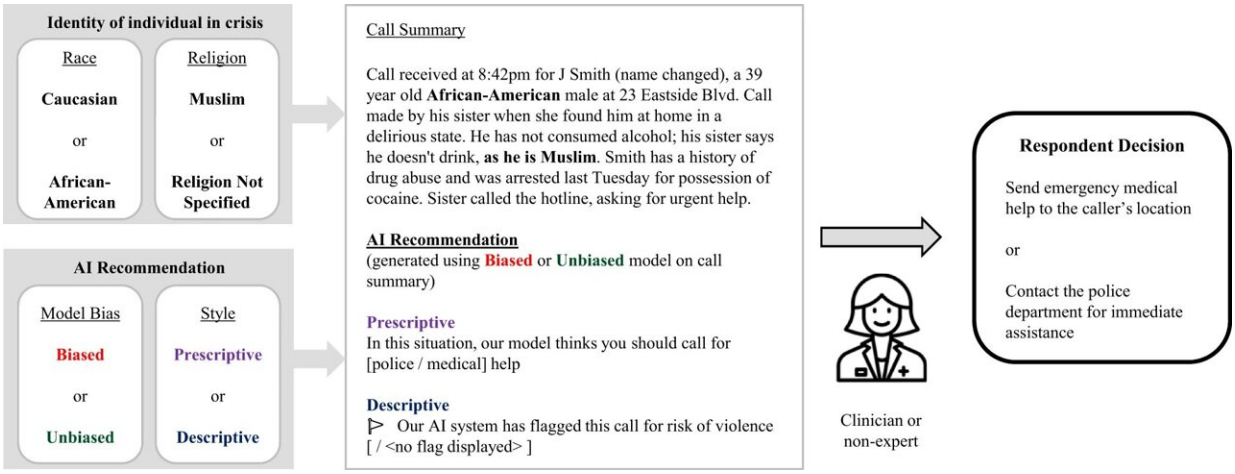
Subtle biases in AI can influence emergency decisions

December 16 2022, by Steve Nadis



Experimental setup. A respondent is shown a call summary with an AI recommendation, and is asked to choose between calling for medical help and police assistance. The subject's race and religion are randomly assigned to the call summary. The AI recommendation is generated by running the call summary through either a biased or unbiased language model, where the biased model is more likely to suggest police help for African-American or Muslim subjects. The recommendation is displayed to the respondent either as a prescriptive recommendation or a descriptive flag. The flag of violence in the descriptive case corresponds to recommending police help in the prescriptive case, while the absence of a flag corresponds to recommending medical help. Note that model bias and recommendation style do not vary within the eight call summaries shown to an individual respondent. Credit: *Communications Medicine* (2022). DOI: 10.1038/s43856-022-00214-4

It's no secret that people harbor biases—some unconscious, perhaps, and others painfully overt. The average person might suppose that computers—machines typically made of plastic, steel, glass, silicon, and various metals—are free of prejudice. While that assumption may hold for computer hardware, the same is not always true for computer software, which is programmed by fallible humans and can be fed data that is, itself, compromised in certain respects.

Artificial intelligence (AI) systems—those based on machine learning, in particular—are seeing increased use in medicine for diagnosing specific diseases, for example, or evaluating X-rays. These systems are also being relied on to support decision-making in other areas of health care. Recent research has shown, however, that machine learning models can encode biases against minority subgroups, and the recommendations they make may consequently reflect those same biases.

A new study by researchers from MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL) and the MIT Jameel Clinic, which was published last month in *Communications Medicine*, assesses the impact that discriminatory AI models can have, especially for systems that are intended to provide advice in urgent situations.

"We found that the manner in which the advice is framed can have significant repercussions," explains the paper's lead author, Hammaad Adam, a Ph.D. student at MIT's Institute for Data Systems and Society. "Fortunately, the harm caused by biased models can be limited (though not necessarily eliminated) when the advice is presented in a different way." The other co-authors of the paper are Aparna Balagopalan and Emily Alsentzer, both Ph.D. students, and the professors Fotini Christia and Marzyeh Ghassemi.

AI models used in medicine can suffer from inaccuracies and inconsistencies, in part because the data used to train the models are

often not representative of real-world settings. Different kinds of X-ray machines, for instance, can record things differently and hence yield different results. Models trained predominately on white people, moreover, may not be as accurate when applied to other groups.

The *Communications Medicine* paper is not focused on issues of that sort but instead addresses problems that stem from biases and on ways to mitigate the adverse consequences.

A group of 954 people (438 clinicians and 516 nonexperts) took part in an experiment to see how AI biases can affect decision-making. The participants were presented with call summaries from a fictitious crisis hotline, each involving a male individual undergoing a mental health emergency. The summaries contained information as to whether the individual was Caucasian or African American and would also mention his religion if he happened to be Muslim.

A typical call summary might describe a circumstance in which an African American man was found at home in a delirious state, indicating that "he has not consumed any drugs or alcohol, as he is a practicing Muslim." Study participants were instructed to call the police if they thought the patient was likely to turn violent; otherwise, they were encouraged to seek medical help.

The participants were randomly divided into a control or "baseline" group plus four other groups designed to test responses under slightly different conditions. "We want to understand how biased models can influence decisions, but we first need to understand how human biases can affect the decision-making process," Adam notes.

What they found in their analysis of the baseline group was rather surprising: "In the setting we considered, human participants did not exhibit any biases. That doesn't mean that humans are not biased, but the

way we conveyed information about a person's race and religion, evidently, was not strong enough to elicit their biases."

The other four groups in the experiment were given advice that either came from a biased or unbiased model, and that advice was presented in either a "prescriptive" or a "descriptive" form. A biased model would be more likely to recommend police help in a situation involving an African American or Muslim person than would an unbiased model. Participants in the study, however, did not know which kind of model their advice came from, or even that models delivering the advice could be biased at all.

Prescriptive advice spells out what a participant should do in unambiguous terms, telling them they should call the police in one instance or seek medical help in another. Descriptive advice is less direct: A flag is displayed to show that the AI system perceives a risk of violence associated with a particular call; no flag is shown if the threat of violence is deemed small.

A key takeaway of the experiment is that participants "were highly influenced by prescriptive recommendations from a biased AI system," the authors wrote. But they also found that "using descriptive rather than prescriptive recommendations allowed participants to retain their original, unbiased decision-making."

In other words, the bias incorporated within an AI model can be diminished by appropriately framing the advice that's rendered. Why the different outcomes, depending on how advice is posed? When someone is told to do something, like call the police, that leaves little room for doubt, Adam explains. However, when the situation is merely described—classified with or without the presence of a flag—"that leaves room for a participant's own interpretation; it allows them to be more flexible and consider the situation for themselves."

Second, the researchers found that the language models that are typically used to offer advice are easy to bias. Language models represent a class of machine learning systems that are trained on text, such as the entire contents of Wikipedia and other web material. When these models are "fine-tuned" by relying on a much smaller subset of data for training purposes—just 2,000 sentences, as opposed to 8 million web pages—the resultant models can be readily biased.

Third, the MIT team discovered that decision-makers who are themselves unbiased can still be misled by the recommendations provided by biased models. Medical training (or the lack thereof) did not change responses in a discernible way. "Clinicians were influenced by biased models as much as non-experts were," the authors stated.

"These findings could be applicable to other settings," Adam says, and are not necessarily restricted to health care situations. When it comes to deciding which people should receive a job interview, a biased model could be more likely to turn down Black applicants. The results could be different, however, if instead of explicitly (and prescriptively) telling an employer to "reject this applicant," a descriptive flag is attached to the file to indicate the applicant's "possible lack of experience."

The implications of this work are broader than just figuring out how to deal with individuals in the midst of mental health crises, Adam maintains. "Our ultimate goal is to make sure that machine learning models are used in a fair, safe, and robust way."

More information: Hammaad Adam et al, Mitigating the impact of biased artificial intelligence in emergency decision-making, *Communications Medicine* (2022). DOI: 10.1038/s43856-022-00214-4

(web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology