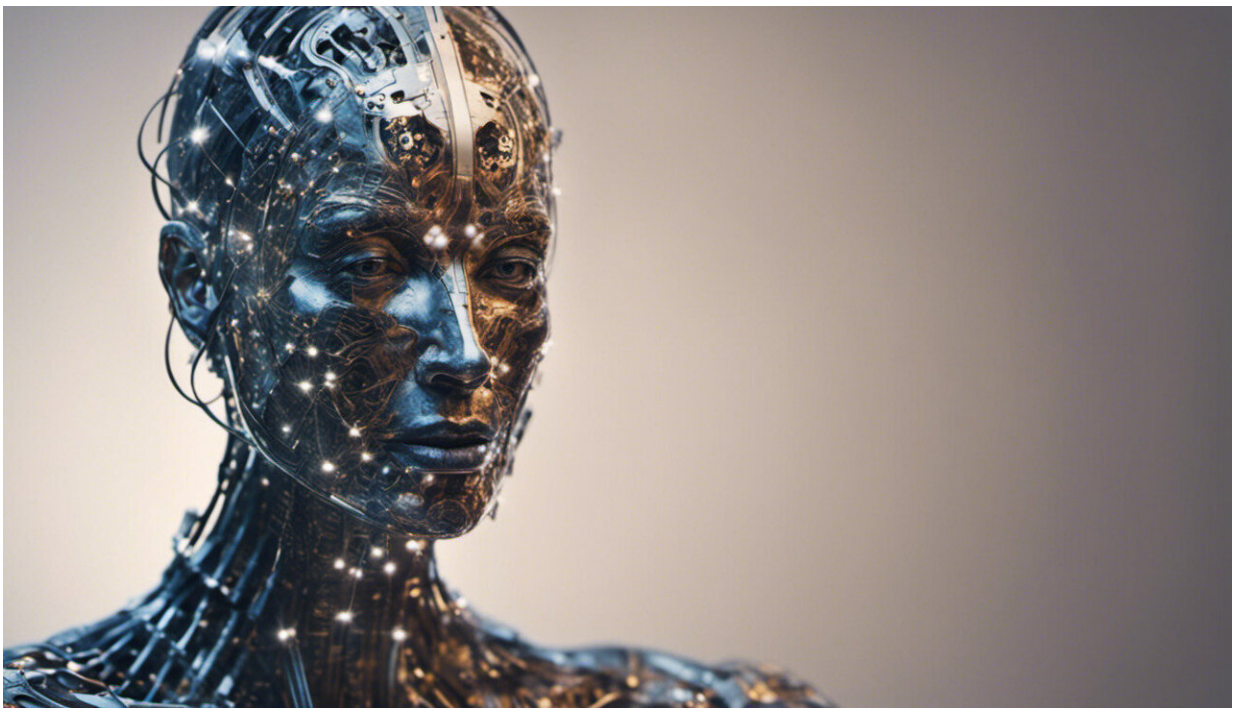
Artificial intelligence must know when to ask for human help

March 7 2019, by Sarah Scheffler, Adam D. Smith And Ran Canetti



Credit: AI-generated image (disclaimer)

Artificial intelligence systems are powerful tools for businesses and governments to process data and respond to changing situations, whether on the stock market or on a battlefield. But there are still some things AI isn't ready for.

We are scholars of computer science working to understand and improve the ways in which algorithms interact with society. AI systems perform best when the goal is clear and there is high-quality data, like when they are asked to distinguish between different faces after learning from many pictures of correctly identified people.

Sometimes AI systems do so well that users and observers are surprised at how perceptive the technology is. However, sometimes success is difficult to measure or defined incorrectly, or the training data does not match the task at hand. In these cases, AI algorithms tend to fail in unpredictable and spectacular ways, though it's not always immediately obvious that something has even gone wrong. As a result, it's important to be wary of the hype and excitement about what AI can do, and not assume the solution it finds is always correct.

When algorithms are at work, there should be a human safety net to prevent harming people. Our research demonstrated that in some situations algorithms can recognize problems in how they're operating, and ask for human help. Specifically, we show, asking for human help can help alleviate algorithmic bias in some settings.

How sure is the algorithm?

Artificial intelligence systems are being used in criminal sentencing, facial-based personality profiling, resume screening, health care enrollment and other difficult tasks where people's lives and well-being are at stake. U.S. government agencies are beginning to ramp up their exploration and use of AI systems, in response to a recent executive order from President Donald Trump.

It's important to remember, though, that AI can cement misconceptions in how a task is addressed, or magnify existing inequalities. This can happen even when no one told the algorithm explicitly to treat anyone

differently.

For instance, many companies have algorithms that try to determine features about a person by their face – say to guess their gender. The systems developed by U.S. companies tend to do significantly better at categorizing white men than they do women and darker-skinned people; they do worst at dark-skinned women. Systems developed in China, however, tend to do worse on white faces.

The difference is not because one group has faces that are easier to classify than others. Rather, both algorithms are typically trained on a large collection of data that's not as diverse as the overall human population. If the data set is dominated by a particular type of face – white men in the U.S., and Chinese faces in China – then the algorithm will probably do better at analyzing those faces than others.

Credit: Unsplash/CC0 Public Domain

No matter how the difference arises, the result is that algorithms can be biased by being more accurate on one group than on another.

Keeping a human eye on AI

For high-stakes situations, the algorithm's confidence in its own result – its estimation of how likely it is that the system came up with the right answer – is just as important as the result itself. The people who receive the output from algorithms need to know how seriously to take the results, rather than assuming that it's correct because it involved a computer.

Only recently have researchers begun to develop ways to identify, much less attempt to fix, inequalities in algorithms and data. Algorithms can be programmed to recognize their own shortcomings – and follow that recognition with a request for a person to assist with the task.

Many types of AI algorithms already calculate an internal confidence level – a prediction of how well it did at analyzing a particular piece of input. In facial analysis, many AI algorithms have lower confidence on darker faces and female faces than for white male faces. It's unclear how much this has been taken into account by law enforcement for high-stakes uses of these algorithms.

The goal is for the AI itself to locate the areas where it is not reaching the same accuracy for different groups. On these inputs, the AI can defer its decision to a human moderator. This technique is especially

well-suited for context-heavy tasks like content moderation.

Human content moderators cannot keep up with the flood of images being posted on social media sites. But AI content moderation is famous for failing to take into account the context behind a post – misidentifying discussions of sexual orientation as explicit content, or identifying the Declaration of Independence as hate speech. This can end up inaccurately censoring one demographic or political group over another.

To get the best of both worlds, our research suggests scoring all content in an automated fashion, using the same AI methods already common today. Then our approach uses newly proposed techniques to automatically locate potential inequalities in the accuracy of the algorithm on different protected groups of people, and to hand over the decisions about certain individuals to a human. As a result, the algorithm can be completely unbiased about those people on which it actually decides. And humans decide on those individuals where algorithmic decision would have inevitably created bias.

This approach does not eliminate bias: It just "concentrates" the potential for bias on a smaller set of decisions, which are then handled by people, using human common sense. The AI can still perform the bulk of the decision-making work.

This is a demonstration of a situation where an AI algorithm working together with a human can reap the benefits and efficiency of the AI's good decisions, without being locked into its bad ones. Humans will then have more time to work on the fuzzy, difficult decisions that are critical to ensuring fairness and equity.

This article is republished from The Conversation under a Creative Commons license. Read the original article.

Provided by The Conversation

Citation: Artificial intelligence must know when to ask for human help (2019, March 7) retrieved 28 June 2024 from https://phys.org/news/2019-03-artificial-intelligence-human.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.