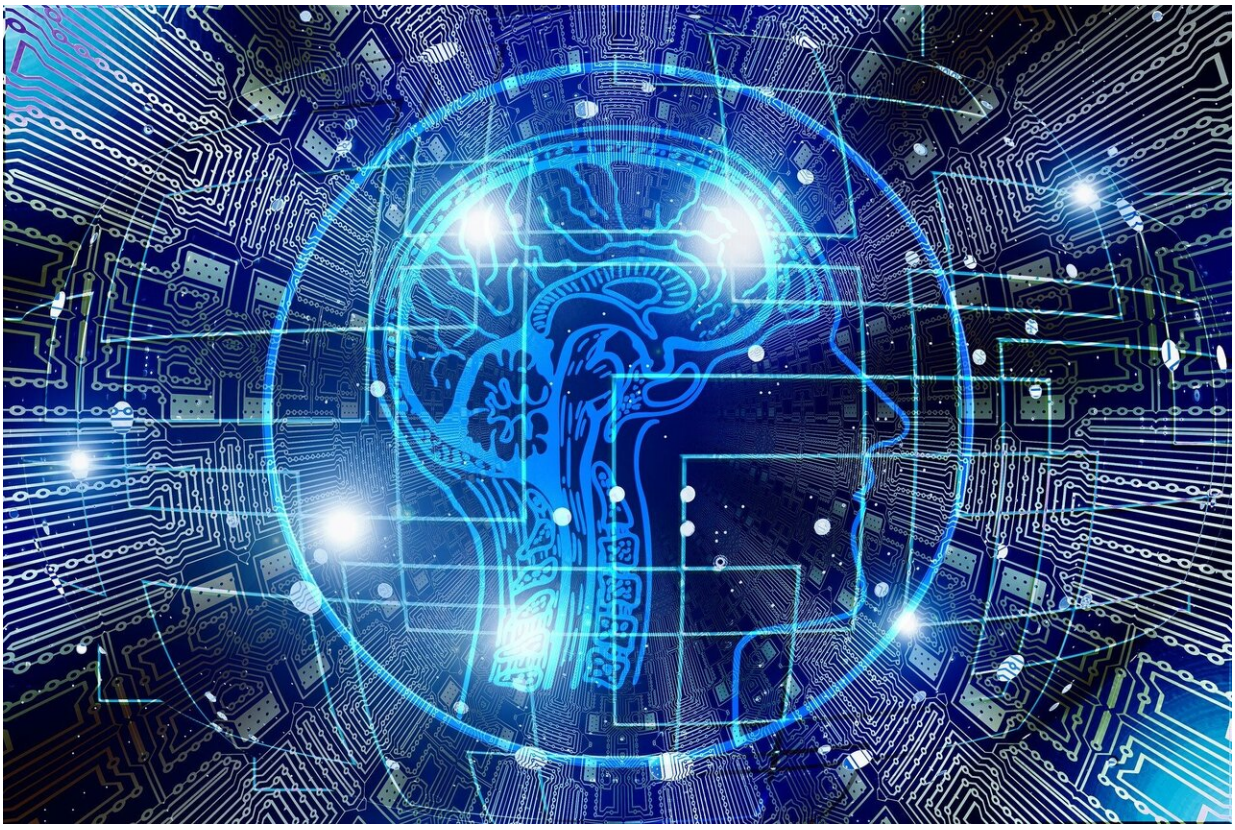
Understanding AI outputs: Study shows pro-western cultural bias in the way AI decisions are explained

April 19 2024



Credit: CC0 Public Domain

Humans are increasingly using artificial intelligence (AI) to inform decisions about our lives. AI is, for instance, helping to make hiring choices and offer medical diagnoses.

If you were affected, you might want an explanation of why an AI system produced the decision it did. Yet AI systems are often so computationally complex that not even their designers fully know how the decisions were produced. That's why the development of "explainable AI" (or XAI) is booming. Explainable AI includes systems that are either themselves simple enough to be fully understood by people, or that produce easily understandable explanations of other, more complex AI models' outputs.

Explainable AI systems help AI engineers to monitor and correct their models' processing. They also help users to make informed decisions about whether to trust or how best to use AI outputs.

Not all AI systems need to be explainable. But in high-stakes domains, we can expect XAI to become widespread. For instance, the recently adopted European AI Act, a forerunner for similar laws worldwide, protects a "right to explanation." Citizens have a right to receive an explanation about an AI decision that affects their other rights.

But what if something like your cultural background affects what explanations you expect from an AI?

In a recent systematic review we analyzed more than 200 studies from the last 10 years (2012–2022) in which the explanations given by XAI systems were tested on people. We wanted to see to what extent researchers indicated awareness of cultural variations that were potentially relevant for designing satisfactory explainable AI.

Our findings suggest that many existing systems may produce explanations that are primarily tailored to individualist, typically western, populations (for instance, people in the U.S. or U.K.). Also, most XAI user studies only sampled western populations, but unwarranted generalizations of results to non-western populations were pervasive.

Cultural differences in explanations

There are two common ways to explain someone's actions. One involves invoking the person's beliefs and desires. This explanation is internalist, focused on what's going on inside someone's head. The other is externalist, citing factors like social norms, rules, or other factors that are outside the person.

To see the difference, think about how we might explain a driver's stopping at a red traffic light. We could say, "They believe that the light is red and don't want to violate any traffic rules, so they decided to stop." This is an internalist explanation. But we could also say, "The lights are red and the traffic rules require that drivers stop at red lights, so the driver stopped." This is an externalist explanation.

Many psychological studies suggest internalist explanations are preferred in "individualistic" countries where people often view themselves as more independent from others. These countries tend to be in the west, educated, industrialized, rich, and democratic.

However, such explanations are not obviously preferred over externalist explanations in "collectivist" societies, such as those commonly found across Africa or south Asia, where people often view themselves as interdependent.

Preferences in explaining behavior are relevant for what a successful XAI output could be. An AI that offers a medical diagnosis might be

accompanied by an explanation such as: "Since your symptoms are fever, sore throat and headache, the classifier *thinks* you have flu." This is internalist because the explanation invokes an "internal" state of the AI—what it "thinks"—albeit metaphorically. Alternatively, the diagnosis could be accompanied by an explanation that does not mention an internal state, such as: "Since your symptoms are fever, sore throat and headache, based on its training on diagnostic inclusion criteria, the classifier produces the output that you have flu." This is externalist. The explanation draws on "external" factors like inclusion criteria, similar to how we might explain stopping at a traffic light by appealing to the rules of the road.

If people from different cultures prefer different kinds of explanations, this matters for designing inclusive systems of explainable AI.

Our research, however, suggests that XAI developers are not sensitive to potential cultural differences in explanation preferences.

Overlooking cultural differences

A striking 93.7% of the studies we reviewed did not indicate awareness of cultural variations potentially relevant to designing explainable AI. Moreover, when we checked the cultural background of the people tested in the studies, we found 48.1% of the studies did not report on cultural background at all. This suggests that researchers did not consider cultural background to be a factor that could influence the generalizability of results.

Of those that did report on cultural background, 81.3% only sampled western, industrialized, educated, rich and democratic populations. A mere 8.4% sampled non-western populations and 10.3% sampled mixed populations.

Sampling only one kind of population need not be a problem if conclusions are limited to that population, or researchers give reasons to think other populations are similar. Yet, out of the studies that reported on cultural background, 70.1% extended their conclusions beyond the study population—to users, people, humans in general—and most studies did not contain evidence of reflection on cultural similarity.

To see how deep the oversight of culture runs in explainable AI research, we added a systematic "meta" review of 34 existing literature reviews of the field. Surprisingly, only two reviews commented on western-skewed sampling in user research, and only one review mentioned overgeneralizations of XAI study findings.

This is problematic.

Why the results matter

If findings about explainable AI systems only hold for one kind of population, these systems may not meet the explanatory requirements of other people affected by or using them. This can diminish trust in AI. When AI systems make high-stakes decisions but don't give you a satisfactory explanation, you'll likely distrust them even if their decisions (such as medical diagnoses) are accurate and important for you.

To address this cultural bias in XAI, developers and psychologists should collaborate to test for relevant cultural differences. We also recommend that cultural backgrounds of samples be reported with XAI user study findings.

Researchers should state whether their study sample represents a wider population. They may also use qualifiers like "U.S. users" or "western participants" in reporting their findings.

As AI is being used worldwide to make important decisions, systems must provide explanations that people from different cultures find acceptable. As it stands, large populations who could benefit from the potential of explainable AI risk being overlooked in XAI research.

Provided by The Conversation