Q&A: Alexa, am I happy? How AI emotion recognition falls short

December 19 2023, by Jade McClain

Credit: Pixabay/CC0 Public Domain

Is the fear of public speaking the same as being chased by a bear? Does raising an eyebrow convey amusement or confusion? In 1995, Rosalind Picard, a scientist and inventor, introduced the idea of computers developing the ability to recognize emotions in her book, "Affective Computing."

For the past several years, systems using artificial intelligence have been "learning" to detect and distinguish human emotion by associating feelings such as anger, happiness, and fear, with facial and bodily movements, words, and tone of voice. But are these systems capable of understanding the nuances that differentiate between a smile and a smirk? Do they know that a smile can accompany anger?

Experts such as Steinhardt Assistant Professor Edward B. Kang warn that the answer is no. Kang, author of the research paper "On the Praxes and Politics of AI Speech Emotion Recognition" published in the *2023 ACM Conference on Fairness, Accountability, and Transparency*, writes that speech emotion recognition (SER) is "a technology founded on tenuous assumptions around the science of emotion that not only render it technologically deficient but also socially pernicious."

Along with other critiques, he suggests that current systems are creating a caricatured version of humanity and exclude those such as people with autism who may emote in ways not understood by these systems.

To better understand those shortcomings and their implications for call centers, dating apps, and more, NYU News spoke with Kang about how AI speech emotion recognition works—and doesn't.

How are AI systems learning to detect emotions?

We need to first ask what we mean by emotion. The reality is that there's no scientific consensus on what emotion actually denotes. Are we referring to a personal experience? A physiological response? A set of brain modes? A subjective feeling? Or any combination of these? The most accurate answer is that we don't really know.

Emotion might be a useful, perhaps even a simple "everyday" term, but scientifically, it's a messy one. We all know that a smile doesn't always

mean we're happy. Researchers have long argued that labels such as "fear," "happiness," "sadness," "anger," "surprise," and "disgust" that we use to refer to emotional experiences are fluid and impossible to pin down according to a bounded set of features.

The problem is that given the structure of machine learning, which refers to the statistical techniques that enable so-called AI systems to "work," emotion needs to be bounded and defined concretely, and its measurability must also be conceived along these observable features.

For the construction of emotion detecting AI systems and the datasets that underlie them, this has traditionally involved hiring human actors to perform certain facial expressions or vocalizations meant to stereotypically represent certain emotional labels—for instance, smiling for "happiness" or shouting for "anger."

These performances become proxies for emotion, writ large, which allows for statistical correlations between observable features such as the tone and speed of one's voice, and the intended "emotion," defined by a "label," to be made. As one can imagine, this results in caricatures of arguably one of the most complex features of humanity.

What are the limitations and harms associated with these systems? What are the benefits?

The limitations of emotion recognition AI systems are that they're by design dependent on the simplification of whatever it is we are defining as emotion in the dataset. In other words, they're just not very reliable or accurate. The harms are that they can still be used as a form of affective surveillance.

As part of my research, I examined the use of speech emotion

recognition in call centers. Here, call center operators are evaluated on whether they sound sufficiently pleasant or not. If they're evaluated positively enough, they can receive compensation bonuses. The flipside, of course, are presumably penalties for not adhering to the emotional norms enforced by the SER system.

Although AI systems are dependent on the thesis that objective emotional definitions exist, the datasets upon which they are trained reveal otherwise. These datasets are ultimately constructed according to the beliefs of the dataset creators and the actors that are hired to perform the emotions—subjective and arbitrary processes through which a few individuals define and perform emotion. These interpretations of emotion get solidified as ground truth in these AI systems.

The benefits of these systems only exist for those who aren't subject to its evaluations. It offers managers, for instance, an additional tool and datapoint for employee evaluation. Even though that data point might not necessarily be what it represents, it offers a level of control for those who use it to evaluate others.

What technologies are currently using and implementing AI speech emotion recognition?

Outside of their application in call centers, AI SER and SER-adjacent voice analytics technologies are being proposed as solutions for higher-stakes contexts such as in finance with loan default prediction, recruiting with candidate success prediction, and the medical field with mental health screenings. To my knowledge, it hasn't been widely implemented yet across these other sectors, but that's also why this is the time to be talking about it.

Microsoft has already committed to removing facial emotion recognition

features from its facial recognition technologies for the same reasons that I draw upon to critique SER, which is that there's a lacking scientific consensus on whether AI-assisted emotion recognition can be done in a way that's reliable, accurate, or consistent. This makes it especially concerning that SER may emerge as a potential replacement for facial emotion recognition.

Based on interviews I've done with industry practitioners, it appears SER is also being proposed for dating apps, which would purportedly aid in providing better matches between individuals.

What are your recommendations for incorporating emotion recognition into consumer products?

My personal recommendation is honestly to not do it at all. In my opinion, it is at best an opt-in "fun" feature for low-stakes applications such as self-monitoring apps, and if it's incorporated as such, it should be made clear that it's for enjoyment purposes only. At its worst, I believe emotion recognition AI is a technological application of a scientifically contentious topic that's used to make life-altering decisions for people who have little to no control over the development and use of these systems.

Affective surveillance and the consequences for compensation examined in the use of SER in call centers is only the beginning of how it can be abused once we accept the problematic premise that emotion can be neatly distilled into data and that a data infrastructure, or what we call "AI," can be leveraged to reliably, accurately, and consistently recognize emotion.

Do you have any thoughts on toys that use SER for interacting with children?

One application that comes to mind is a toy robot called Moxie that incorporates multimodal AI emotion recognition in its engagement with children. Based on a paper released by its creators, the behavioral metrics that the toy tracks primarily relate to facial expressions and word choices. Here, even though the word choices are technically recorded via speech through a microphone, it's different from SER because the analysis of words is presumably powered first by a speech-to-text model that converts the speech into text, and then analyzes that text to examine if certain words, such as "family" or "friend," relate to concepts that they deem to be "positive" or "negative."

This is generally called "sentiment analysis" in the field, and it's also a somewhat contentious area for similar reasons: words alone are not consistently indicative of "sentiment." The paper states that the toy was first developed as a tool for supporting children diagnosed with mental behavioral development disorders or MBDDs, but my understanding is that it's now being sold as a more general learning companion for all children that supports "holistic skill development," which of course expands the addressable market of Moxie.

My colleague Mara Mills has called this phenomenon of resourcing disability as a step towards more profitable realms as "assistive pretext." As I briefly recount in my paper, children, and especially those who have been diagnosed with MBDDs, have historically been designated as the target demographic and justification for the initial development of emotion recognition technologies.

A chapter in Rosalind Picard's pioneering 1995 book "Affective Computing," for instance, has a section dedicated to "helping autistic individuals." About a decade later, researchers from the University of Cambridge also proposed an "emotional hearing aid" that was described as a facial prosthetic to help children with Asperger's Syndrome socialize. To my knowledge, most of this work as it has been taken up by

the broader tech industry has now developed beyond these "assistive pretexts," and the benefit for the individuals that served as the justification for their initial development is contestable. My hope is for researchers and builders to remain critical and compassionate in their development, or not, of these technologies.

More information: Edward B. Kang, On the Praxes and Politics of AI Speech Emotion Recognition, *2023 ACM Conference on Fairness, Accountability, and Transparency* (2023). DOI: 10.1145/3593013.3594011

Provided by New York University