Break The Rules? AI Will Show You No Mercy

Description

AI is emerging as the judge, jury and executioner. If you break the rules, there will be no mercy or human understanding. AI is incapable of human understanding or emotions. Punishment will be swift and there will be no appeal or other recourse to clear your name. Science fiction has dealt with this on several occasions, such as RoboCop, Judge Dredd and the Minority Report. ? TN Editor

You might think a computer would be an unbiased and fair judge, but a new study finds you might be better leaving your fate in the hands of humans. Researchers from MIT find that artificial intelligence (AI) tends to make stricter and harsher judgments than humans when it comes to people who violate the rules. Simply put, AI isn't willing to let people off the hook easy when they break the law!

Researchers have expressed concerns that AI might impose overly severe punishments, depending on the information scientists program it with. When AI is programmed strictly based on rules, devoid of any human nuances, it tends to respond harshly compared to when it is programmed based on human responses.

This study, conducted by a team at the Massachusetts Institute of Technology, examined how AI would interpret perceived violations of a given code. They discovered that the most effective data to program AI with is normative data, where humans have determined whether a specific rule has been violated. However, many models are erroneously programmed with descriptive data, in which people label the factual attributes of a situation, and AI determines whether a code has been breached.

In the study, the team gathered images of dogs that could potentially violate an apartment rule banning aggressive breeds from the building. Groups were then asked to provide normative and descriptive responses.

The descriptive team wasn't informed about the overall policy on dogs, and was asked to identify whether three factual elements, such as the dog's aggression, were present in the image or text. Their responses helped to form judgments. If a user said the photo depicted an aggressive dog, the policy was considered violated. On the other hand, the normative group was informed about the rules on aggressive dogs and was asked to determine whether each image violated the rule, and if so, why.

Participants were 20 percent more likely to identify a code violation using the descriptive method compared to the normative one. If the descriptive data on dog behavior had been used to program an AI model, it would be more likely to issue severe penalties.

Scaling up these inaccuracies to real-world scenarios could have substantial implications. For instance, if a descriptive model is used to predict whether a person may commit the same crime more than once, it may impose harsher judgments than a human and result in higher bail amounts or longer criminal sentences. Consequently, the experts have advocated for increased data transparency, arguing that understanding how data is collected can help determine its potential uses.

"Most AI/machine-learning researchers assume that human judgments in data and labels are biased. But our results indicate a more troubling issue: these models are not even reproducing already-biased human judgments because the data they're being trained on is flawed," says Marzyeh Ghassemi, an assistant professor and head of the Healthy ML Group in the Computer Science and Artificial Intelligence Laboratory (CSAIL), in a university release.

"The solution is to acknowledge that if we want to reproduce human judgment, we should only use data collected in that context. Otherwise, we'll end up with systems that impose extremely harsh moderations, far stricter than what humans would impose. Humans would see nuances or make distinctions, whereas these models don't," Ghassemi further explains.

In the study, published in *Science Advances*, the team tested three additional datasets. The results varied, ranging from an eight-percent increased likelihood of identifying a rule violation using descriptive responses for a dress code violation, up to a 20-percent increase for the aggressive dog images.

"Perhaps the way people think about rule violations differs from how they think about descriptive data. Generally, normative decisions tend to be more lenient," says lead author Aparna Balagopalan. "The data really matter. It's crucial to align the training context with the deployment context when training models to detect rule violations."

The team's future plan is to investigate the impact of having professionals, such as lawyers and doctors, participate in data entry.

Category

1. Main
2. NWO-Deep State-Dictatorship-Tyrrany
3. Science-Tech-AI-Medical & Gen. Research

Date Created
05/17/2023