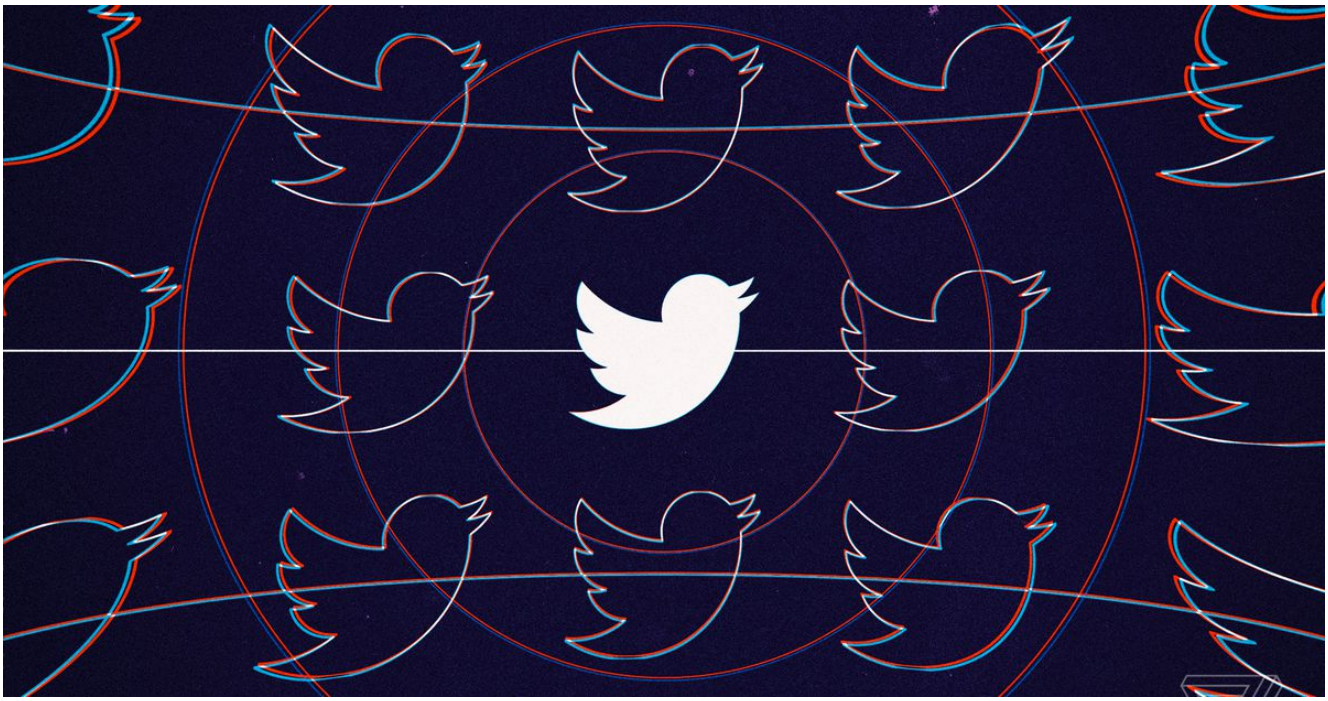
After accusations, Twitter will pay hackers to find biases in its automatic image crops



Twitter is holding a competition in hopes that hackers and researchers will be able to identify biases in its image cropping algorithm — and it's going to be handing out cash prizes to winning teams (via *Engadget*). Twitter is hoping that giving teams access to its code and image cropping model will let them find ways that the algorithm could be harmful (such as it cropping in a way that stereotypes or erases the image's subject).

Those competing will have to submit a description of their findings, and a dataset that can be run through the algorithm to demonstrate the issue. Twitter will then assign points based on what kind of harms are found, how much it could potentially affect people, and more.

The winning team will be awarded $3,500, and there are

separate $1,000 prizes for the most innovative and most generalizable findings. That amount has caused a bit of a stir on Twitter, with a few users saying it should have an extra zero. For context, Twitter's normal bug bounty program would pay you $2,940 if you found a bug that let you perform actions for someone else (like retweeting a tweet or image) using cross-site scripting. Finding an OAuth issue that lets you take over someone's Twitter account would net you $7,700.

A competition lets Twitter get feedback from a much broader range of perspectives

Twitter has done its own research into its image-cropping algorithm before — in May, it published a paper investigating how the algorithm was biased, after accusations that its previews crops were racist. Twitter's mostly done away with algorithmically cropping previews since then, but it's still used on desktop and a good cropping algorithm is a handy thing for a company like Twitter to have.

Opening up a competition lets Twitter get feedback from a much broader range of perspectives. For example, the Twitter team held a space to discuss the competition during which a team member mentioned getting questions about caste-based biases in the algorithm, something that may not be noticeable to software developers in California.

Twitter is also looking for ways its algorithm could be exploited

It's also not just subconscious algorithmic bias Twitter is looking for. The rubric has point values for both intentional and unintentional harms. Twitter defines unintentional harms as crops that could result from a "well-intentioned" user posting a regular image on the platform, whereas intentional harms are problematic cropping behaviors that could be exploited by someone posting maliciously designed images.

Twitter says in its announcement blog that the competition is separate from its bug bounty program — if you submit a report

about algorithmic biases to Twitter outside of the competition, the company says your report will be closed and marked as not applicable. If you’re interested in joining, you can head over to the competition’s HackerOne page to see the rules, criteria, and more. Submissions are open until August 6th at 11:59PM PT, and the winners of the challenge will be announced at the Def Con AI Village on August 9th.

https://www.theverge.com/2021/7/30/22602553/twitter-image-cropping-algorithm-bias-competition-bounty-def-con