to the U.S.?

A. 50% B. 20% C. 78% D. 90% E. 15%

from Gartner Inc. citing the figures. The additional spending is being spurred in part by

Cybersecurity, also contributed to this quiz. She can be reached at *kim.nash@wsj.com*.

Attackers Devise Ways to Get Around Image-Recognition Filters

By Parmy Olson

Last year, engineers at **ZeroFOX**, a security startup, noticed something odd about a fake social-media profile they'd found of a well-known public figure. Its profile photo had tiny white dots across the face, like a dusting of digital snow. The company's engineers weren't certain, but it looked like the dots were placed to trick a content filter, the kind used by social networks like Facebook to flag celebrity imitations.

They believed the photo was an example of a new kind of digital camouflage, sometimes called an adversarial attack, in which a picture is altered in ways that leave it looking normal to the human eye but cause an image-recognition system to misclassify the image.

Such tricks could pose a security risk in the global rush among businesses and governments to use image-recognition technology. In addition to its use in social-network filters, image-recognition software shows up in security systems, self-driving cars and many other places, and tricks like this underscore the challenge of keeping such systems from being fooled or gamed.

One senior technology executive says groups of online attackers have been launching "probing attacks" on the content filters of social-media companies. Those companies have ramped up their efforts to eliminate banned content—everything from child pornography and terrorist messages to fake profiles—with expanded content filters.

"There's a bunch of work on attacking AI algorithms, changing a few pixels," the executive says. "There have been groups trying to use these attacks on some of the large social-media companies in the U.S."

A spokesman for **Facebook** said the company was aware of users trying to trick its image-recognition systems, a technique it refers to internally as "image and video content matching." Such users were often trying to sell banned items like drugs or guns in Facebook groups or on ads, but most approaches were rudimentary, the spokesman said. Some users, for example, tried to bypass filters by using photos of cannabis that looked like fried broccoli; Facebook correctly flagged it. The spokesman said he wasn't aware of more-sophisticated attempts to digitally disguise an image, and emphasized Facebook was mostly blocking fake accounts and spam, while guns, nude pictures and drugs were a minor portion of banned content. "Those are several orders of magnitude smaller," he said.

Researchers found that a special colorful poster rendered the person holding it invisible to image-classification software.

Facebook struggled to handle another low-tech form of adversarial attack in April, when millions of copies of the live-streamed video of the gunman who killed 51 people in two mosques in Christchurch, New Zealand, kept getting uploaded to the site. Facebook blamed a "core community of bad actors." Their methods were rudimentary and involved slightly editing the videos or filming them and re-uploading new copies, so that Facebook couldn't rely on the digital fingerprint it had assigned the initial video. Facebook also struggled because its image-recognition system for flagging terrorist content had been trained on videos filmed by a third person, not a first-person perspective the gunman had used, the spokesman said.

Facebook has expanded its use of artificial intelligence in recent years. While the company has hired 30,000 human content moderators, it relies primarily on artificial intelligence to flag or remove hate speech, terrorist propaganda and spoofed accounts. Image recognition is one form of artificial intelligence typically used to screen the content that people post, because it can identify things like faces, objects or a type of activity.

Google has said it also plans to increasingly rely on using AI-powered software to block toxic content on YouTube. It has hired 10,000 people to help moderate content, but wants that tally of human workers to go down, according to a senior official from the company. "AI solves that problem," the official said. Google declined to comment on whether the company has experienced any adversarial attacks that involved digitally altered images, but pointed to research papers it published last year on how to defend online systems from such attacks.

But a growing body of science shows image-recognition systems' vulnerability to adversarial attacks. One example comes from an experiment from September 2018, where academics took a digital photo of crack cocaine being heated up in a spoon and slightly modified its pixels. The image became a little fuzzier to humans, but was now classified as "safe" by the image-recognition system of **Clarifai** Inc.

Clarifai is a New York-based content-moderation service used by several large online services.

Clarifai said its engineers were aware of the study, but declined to comment on whether it had updated its image-classification system as a result.

"We openly invite both AI researchers and our customers to collaborate with Clarifai to share their findings and conceive defenses against unintended uses of AI models," a spokesman said.

Researchers also have shown that image-recognition systems can be fooled offline.

In April, researchers at KU Leuven, a university in Belgium, tricked a popular image-classification system by holding a small, colorful poster, about the size of a vinyl-record album cover, in front of them while standing before a surveillance camera. The special poster made the person holding it invisible to the software.

Ms. Olson is a Wall Street Journal reporter in London. She can be reached at parmy.olson@wsj.com.

TOP: JON KRAUSE; WIEBE VAN RANST; BOTTOM: SIMEN THYS, WIEBE VAN RANST AND TOON GOEDEMÉ FROM KU LEUVEN