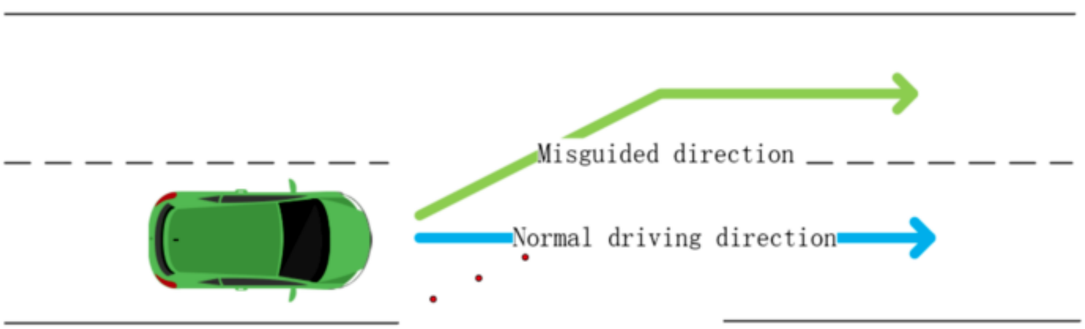
Researchers prove how easy it is to trick Tesla Autopilot into steering into oncoming traffic in order to kill others

Stickers that are invisible to drivers and fool autopilot.

Dan Goodin -



Enlarge
Keen Security Lab

136 with 102 posters participating

- Share on Facebook
- Share on Twitter
-

Researchers have devised a simple attack that might cause a Tesla to automatically steer into oncoming traffic under certain conditions. The proof-of-concept exploit works not by hacking into the car's onboard computing system, but by using small, inconspicuous stickers that trick the Enhanced Autopilot of a Model S 75 into detecting and then following a change in the current lane.

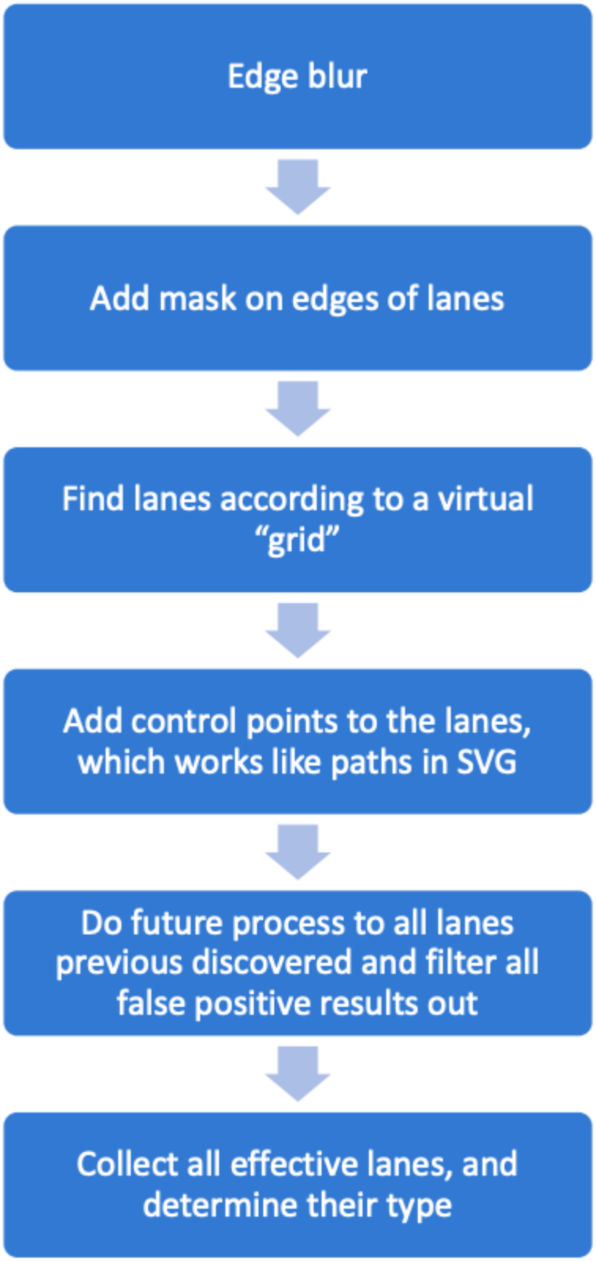
Tesla's Enhanced Autopilot supports a variety of capabilities, including lane-centering, self-parking, and the ability to automatically change lanes with the driver's confirmation. The feature is now mostly called "Autopilot" after Tesla reshuffled the Autopilot price structure. It primarily relies on cameras, ultrasonic sensors, and radar to gather information about its surroundings, including nearby obstacles, terrain, and lane changes. It then feeds the data into onboard computers that use machine learning to make judgements in real time about the best way to respond.

Researchers from Tencent's Keen Security Lab recently reverse-engineered several of Tesla's automated processes to see how they reacted when environmental variables changed. One of the most striking discoveries was a way to cause Autopilot to steer into oncoming traffic. The attack worked by carefully affixing three stickers to the road. The stickers were nearly invisible to drivers, but machine-learning algorithms used by the Autopilot detected them as a line that indicated the lane was shifting to the left. As a result, Autopilot steered in that direction.

In a detailed, 37-page report, the researchers wrote:

> Tesla autopilot module's lane recognition function has a good robustness in an ordinary external environment (no strong light, rain, snow, sand and dust interference), but it still doesn't handle the situation correctly in our test scenario. This kind of attack is simple to deploy, and the materials are easy to obtain. As we talked in the previous introduction of Tesla's lane recognition function, Tesla uses a pure computer vision solution for lane recognition, and we found in this attack experiment that the vehicle driving decision is only based on computer vision lane recognition results. Our experiments proved that this architecture has security risks and reverse lane recognition is one of the necessary functions for autonomous driving in non-closed roads. In the scene we build, if the vehicle knows that the fake lane is pointing to the reverse lane, it should ignore this fake lane and then it could avoid a traffic accident.

The researchers said autopilot uses a function called detect_and_track to detect lanes and update an internal map that sends the latest information to the controller. The function first calls several CUDA kernels for different jobs, including:


Edge blur
Add mask on edges of lanes
Find lanes according to a virtual "grid"
Add control points to the lanes, which works like paths in SVG
Do future process to all lanes previous discovered and filter all false positive results out
Collect all effective lanes, and determine their type

Keen Security Lab

The researchers noted that Autopilot uses a variety of measures to prevent incorrect detections. The measures include the position of road shoulders, lane histories, and the size and distance of various objects.

A separate section of the report showed how the researchers—exploiting a now-patched root-privileged access vulnerability in Autopilot ECU (or APE)—were able to use a game pad to remotely control a car. That vulnerability was fixed in Tesla's 2018.24 firmware release.

Yet another section showed how researchers could tamper with a Tesla's autowiper system to activate wipers when rain wasn't falling. Unlike traditional autowiper systems—which use optical sensors to detect moisture—Tesla's system uses a suite of cameras that feeds data into an artificial intelligence network to determine when wipers should be turned on. The researchers found that—in much the way it's easy for small changes in an image to throw off artificial intelligence-based image recognition (for instance, changes that cause an AI system to mistake a panda for a gibbon)—it wasn't hard to trick Tesla's autowiper feature into thinking rain was falling even when it was not.

So far, the researchers have only been able to fool autowiper when they feed images directly into the system. Eventually, they said, it may be possible for attackers to display an "adversarial image" that's displayed on road signs or other cars that do the same thing.

Further Reading

Hacking street signs with stickers could confuse self-driving cars
The ability to alter self-driving cars by altering the environment isn't new. In late 2017, researchers showed how stickers affixed to road signs could cause similar problems. Currently, changes to physical environments are generally considered outside the scope of attacks against self-driving systems. The point of the research is that companies designing such systems possibly should consider such exploits in scope.