Adversarial Attacks on Autonomous Cars

Imagine cruising down an empty road in your vehicle on its autonomous mode. As you near the a stop sign, your vehicle begins to turn in an odd manner. You suddenly find yourself in the midst of a hasty U-Turn, even though there is clearly a stop sign ahead! You then realize that an adversarial attack is the only way this could have happened.

What are adversarial attacks?

“Adversarial machine learning: is a technique employed in the field of machine learning which attempts to fool models through malicious input. This technique can be applied for a variety of reasons, the most common being to attack or cause a malfunction in standard machine learning models.” - Pin-Yu Chen, AI Researcher at the IBM research facility.

Generally, an AI model aims to perform a task with maximum efficiency and accuracy and with a minimum loss function. However, an attacker would aim to alter these preferred goals by making the AI model perform the task with a maximum loss function and minimum efficiency and accuracy.

How do autonomous vehicles work?

Radar and LiDAR Systems

They both essentially perform the same task: autonomous cars create and maintain a map of their surroundings based on a variety of sensors situated in different parts of the vehicle. However, they execute these tasks in different ways. Additionally, a vast majority of the vehicles autonomous system depends *entirely* upon the accuracy of these sensor systems, be it radar, liDAR or cameras.

Radar sensors monitor the positions of surrounding vehicles. These sensors feed the vehicle control systems with data on the vehicle's surroundings (distancing to the curb of a road, the distance between 2 cars, etc.). The radar system works very similarly to that of liDAR, while talking about principles. Both employ transmitters that shoot out waves in predetermined directions. When these waves contact an object, they are usually reflected and scattered in many directions (some can be absorbed by the objects). Transmitters in radar systems emit radio waves to detect objects around the vehicle.

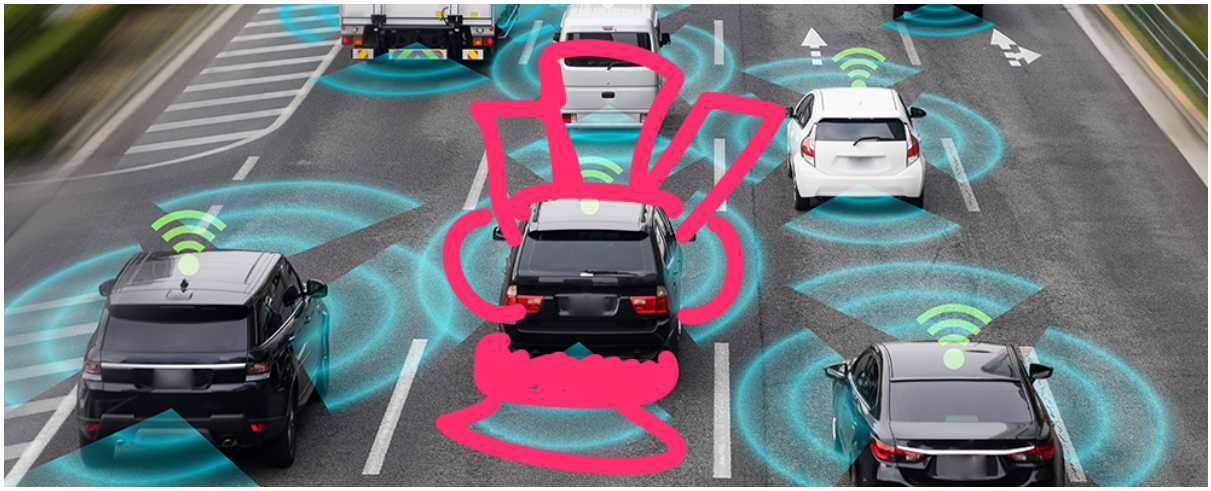


Figure 1 - The radar system in the autonomous cars above are sending radio waves and learning their surroundings

Transmitters in liDAR systems, on the other hand, bounce pulses of light off the car's surroundings to measure distances, detect road edges and identify lane markings. Briefly, liDAR is extremely similar to radar in the sense that it uses principles of echolocation

Figure 2 - The liDAR systems of the autonomous car are sending pulses of light to learn their surroundings

Camera Systems

Autonomous softwares can also use cameras to classify objects around them and detect their surroundings. This is done by using an AI image classifier model in order to detect the type of object with a certain percentage of accuracy. Below is an image of the autonomous car classifying these images using its AI model. One thing to note is that camera systems need to

have a very high efficiency and very low loss function. The faster the vehicle's model is able to detect what the object is, the quicker it can perform the appropriate task in response.

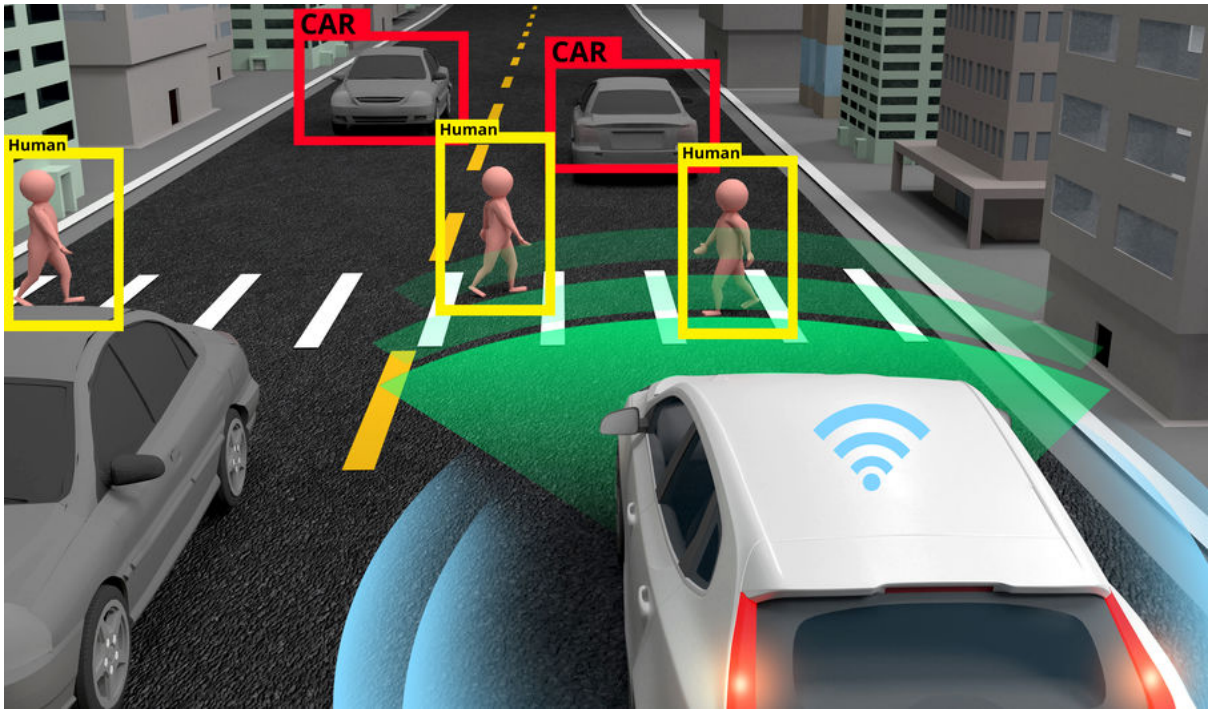


Figure 3 - The autonomous vehicle is detecting objects in its surroundings using cameras, liDAR, and radar systems

How can these autonomous systems be exploited?

Each autonomous model can be exploited in many different ways, but this paper will briefly cover how adversarial attacks can use the weaknesses of these systems to an attackers advantage. Now, an autonomous vehicle can employ radar, liDAR, and/or camera systems. For a vehicle that uses all three (or more) systems, the system would be that much more fail proof.

Say an attacker attempted an adversarial attack on a car that used all three (or more) systems. If the attacker did not know that multiple autonomous systems were guiding the car and they attacked only one system. The rest of the systems would return the same information, whereas the affected system's results would be skewed. However, the software that controls the car would use majority, because there is always a small loss percentage in these models in the cars.

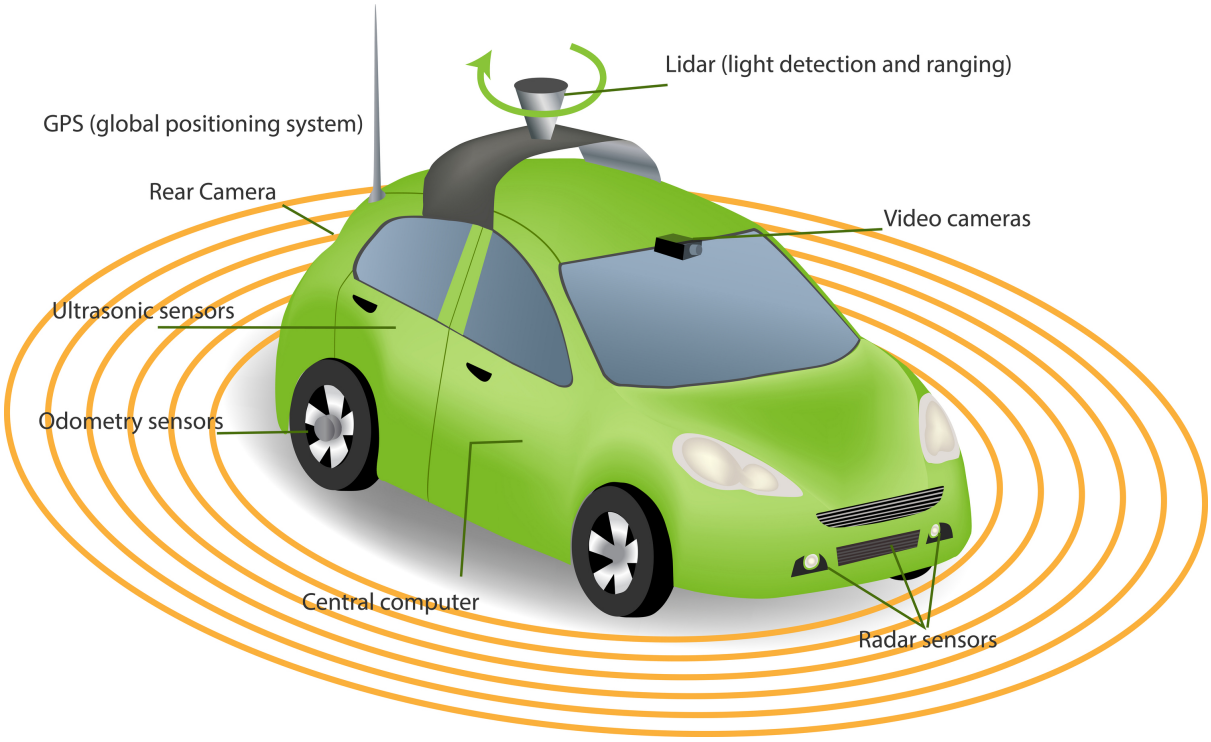


Figure 4 - The image illustrates the possible ways by which an autonomous vehicle can detect its surroundings

How could the radar and liDAR systems be exploited?

Once waves are sent out of the transmitters, be it light waves or radio waves, there is a reader that essentially maps the surroundings of the car in order to determine its next tasks. Let's assume that the wavelengths emitted have a wavelength of 1cm and have a frequency of about 10 Hz. This means that the receiver that 'collects' the waves would have its reading set to the same wavelength of 1cm and the same frequency of 10 Hz. Now, what an attacker can do is alter the parameters of the transmitter in order to confuse the receiver of the car, thereby getting a negative reading.

The receiver would expect to accept the waves of the same wavelength and frequency that were emitted by its transmitter and would ignore waves of all other wavelengths and frequencies. The system would then think that its surroundings would be purely empty because it would believe the waves never returned.

This subtle but effective (for bad reasons) adversarial attack would work with both LiDAR and radar systems.

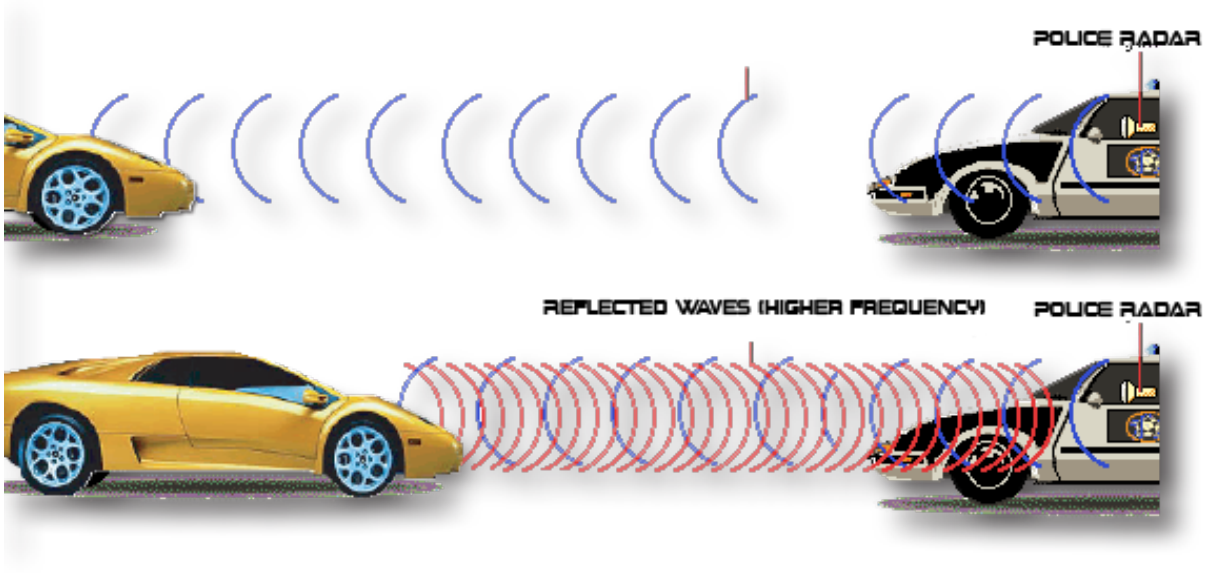


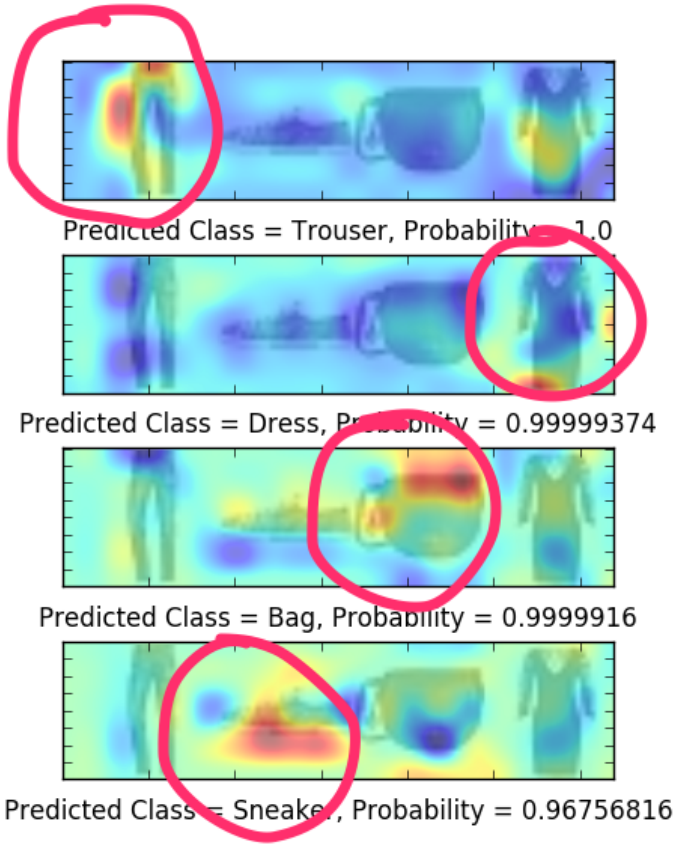
Figure 5 - the yellow car is blocking

Though the image above essentially explains physical liDAR jamming that could be used to prevent police cars from detecting the speed of a vehicle, it provides an interesting and accurate representation of how the adversarial attacks on radar and liDAR systems could work.

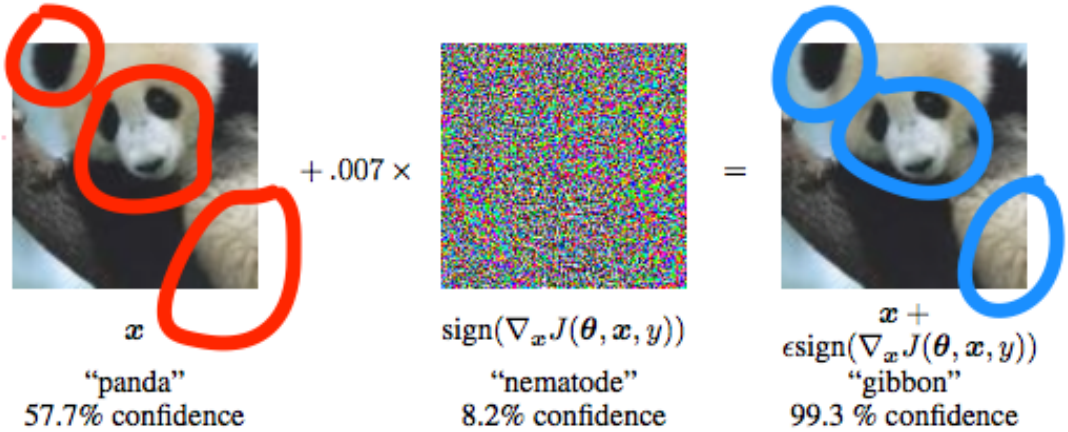
In the image above the red and blue waves are spaced out differently, indicating different wavelengths and/or frequencies. Lets assume the police car on the right emitting the blue waves wants to find out the speed and position of the yellow car using a liDAR speed detector. The yellow car's speed would not register accurately on the police car's system because of the red waves, that have a different wavelength, are being reflected back.

How could a camera system be exploited?

AI models use class activation mapping (CAM) to determine which region of an image is relevant to the certain class in order to identify the discriminative regions, regions in an image that help the AI model determine what it is.



As seen in the image above, the regions circled in red are the discriminative regions of the image. If an attacker adds a certain amount of noise to these discriminative regions, as shown in the picture below, they can confuse the AI model into thinking it is something else.



In the image shown above, noise (overlaid image/audio that causes a classifier to miscategorize an image and is undetectable by the human eye) is added to an image of a panda to fool the AI model into thinking its a gibbon instead. An important thing to note from an image is that this noise is undetectable by the human eye. The regions circled in red are

possible regions that the AI model uses to classify the image as a panda and the regions circled in blue are those possible regions from image *x* with some noise added to it.

If a car is driving down a road and sees a stop sign, the noise added to this image could fool the AI model into thinking that the stop sign is actually a U-turn sign, prompting the car to turn.

How could these adversarial attacks be prevented?

Now, this paper has mainly focused on how these attacks can be harmful, but, in response, there is a lingering question: how could these adversarial attacks be prevented, or in the very least, have its effect reduced significantly to a point where its loss function is as close to a normal functioning model.

For the method of attacking presented earlier, there can be a defense mechanism set up that could *reduce* the loss function significantly as opposed to an attack with no protective measures taken.

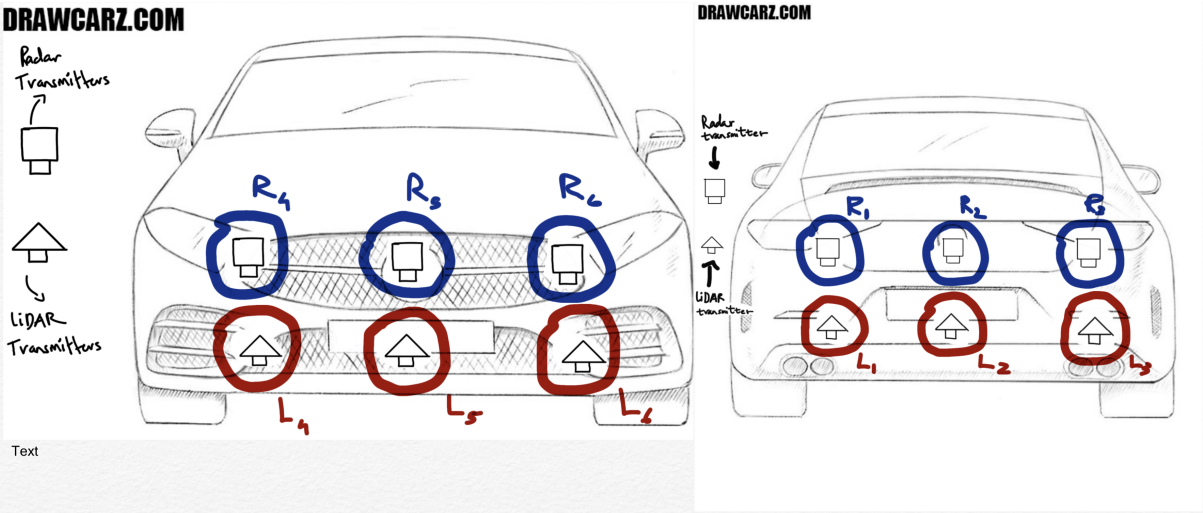


Figure 9 - Front and back of an autonomous car with three radar and liDAR transmitters

Figure 9 Description (same wavelengths for transmitters in front)

- R_1 - wavelength: 1 inch
- R_2 - wavelength: 3 inches
- R_3 - wavelength: 5 inches
- L_1 - wavelength: 1.5 millimeters
- L_2 - wavelength: 2.5 millimeters
- L_3 - wavelength: 3.5 millimeters

In short and simple language, for the liDAR and radar systems, I propose a car with a feedback loop and checking system. In the diagram above, there are 6 transmitters of each type with 3 at each end of the car. Lets assume an attacker writes a program that targets one of the transmitters and makes it malfunction. This way, we have two other transmitters that can fact check the affected transmitter. By realizing that majority of the transmitter data is similar, the program will be able to ignore all data received by the affected transmitter in realization that it has been compromised in some way. Not only would it take an attacker a longer period of time to code, the attacker would have to code a program for different situations as well. This model is very robust and has very low loss function.

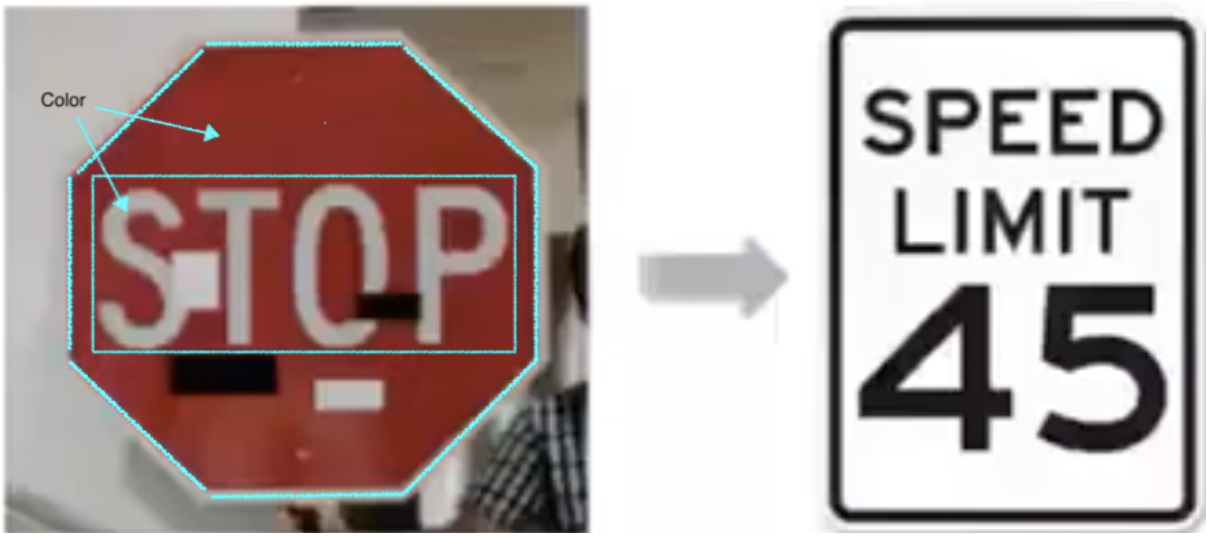
The presence of these transmitters at the front *and* the back of the car allows for an added layer of security and 'vision' for the car to familiarize and recognize its surroundings while navigating in the autonomous mode.

For the camera system, preventing against adversarial attacks get slightly more complicated. This is, of course, for **the proposition I have made** for how the adversarial attacks can be

done on a camera system on the autonomous vehicle.

Now, AI neural network models are trained with millions of images to classify images. I propose a model in which we can train the AI model to recognize images with noise as original image. Let's return to our example of the affected stop sign.

Though recognizable as a 'stop' sign to the human eye, the AI model gets confused and instead reads this image as a speed limit. In my proposition, we should train the AI model with millions of images that include noise as well.



In the image above, the stop sign clearly has some noise added to it. I propose that we focus on these discriminative regions that are outlined in the teal color. The main features of the stop sign are the red color of the stop sign, the white color of the text, the font used and the octagonal shape of the sign. By writing a new code that focuses on these elements as the discriminative regions, the AI model should be able to recognize *any* image, no matter the noise. For example, we should train the AI model to recognize images with noise, like shown above, as stop signs, no matter the amounts of noise.

With the exponential growth of autonomous car companies like Tesla, Waymo, Cruise, and many others, it is important to recognize the potential risks that come along with buying an autonomous vehicle.

Hopefully, this piece has helped you gain a deeper understanding as to what can happen to your autonomous car and how it could pose as a threat to not only the driver, but to the others on the road.