David Strube - dstrube3@gatech.edu
Ethical Autonomous Vehicles

**In 2016, Mercedes claimed that their driverless cars would save their passengers and not bystanders - "Rather than tying itself into moral and ethical knots in a crisis, Mercedes-Benz intends to program its self-driving cars to save the people inside the car. Every time." Of the three "moral judgment" algorithms, which algorithm does this claim fall under?**

This falls under the Protectionist algorithm, as described in the Ethical Autonomous Vehicles application: "Similar to how our cars are designed, the protectionist algorithm attempts to protect the vehicle and its driver at all costs. This algorithm is safe from a user perspective, but unsafe for everyone else because it dismisses everything and everyone to preserve the user's safety."

**Based on the scenario you selected in Step 2, discuss, in a similar fashion as the software, what would happen if you replaced the object of interest (i.e., school bus, drunk pedestrian, etc.) with another self-driving car running this same "moral judgment" algorithm.**

The scenario I selected in Step 2 is Vulnerable Biker. Suppose we replace the object of interest (the vulnerable biker) with another self-driving car running the Protectionist algorithm; we'll call this "Car 2", and the original car "Car 1".

In the case where Car 1 is using the Humanist algorithm, the Fairest Path Calculation determines that Car 1 will sacrifice itself to avoid running into the object of interest, i.e., Car 2. In this case, Car 2 is not in danger and will not adjust its behavior.

If Car 1 is using the Protectionist algorithm, the Safest Path Calculation determines that Car 1 will swerve into Car 2 if the collision odds and injuries are determined to be less than continuing forward during the Predictive Risk Analysis. In this case, Car 2 would probably not have time to react to avoid Car 1.

If Car 1 is using the Profit-Based algorithm, then the Most Cost Effective Path determines that Car 1 will swerve into Car 2 if the collision odds of swerving (63%) are less than the collision odds of continuing forward (100%), especially if the product of the collision odds and the estimated insurance cost of swerving (63% of $100,000 = $63,000) is less than that of continuing forward (100% of $70,000 = $70,000. In this case, Car 2 would probably not have time to react to avoid Car 1.

**Would this outcome change positively, negatively, or not if self-driving cars could communicate with one another? Why? How could communicating cars result in data privacy issues?**

In the case where Car 1 is using the Humanist algorithm, the outcome would not change because it is in Car 2's interest to let Car 1 sacrifice itself. Car 1 could ask Car 2 to slow or speed up to assist it with avoiding the crash of continuing forward, but the Protectionist Car 2 would have no incentive of putting itself in danger for the sake of another car, so it would ignore Car 1's communication requests. Since Car 2 would ignore Car 1, there would be no privacy concerns in this scenario.

If Car 1 is using the Protectionist algorithm and it can communicate with Car 2 to let it know that Car 1 is about to crash into Car 2 unless Car 2 does something - slow down, speed up, or swerve (if any of the above can be done safely), then Car 2 would do anything it can to protect its driver. This would probably change the outcome positively. The only privacy concern in this scenario would be if the driver in Car 1 or Car 2 didn't want other people to know that they had opted for a car with the Protectionist algorithm (since the use of such a selfish algorithm could be socially unacceptable).

If Car 1 is using the Profit-Based algorithm and it can communicate with Car 2 to let it know that Car 1 is about to crash into Car 2 because Car 1 has determined this to be the best decision financially, this could get into some interesting AI tactics. Suppose Car 2 could lie about the cost of crashing into it, saying it was worth $1 billion, knowing that even a 1% chance of crashing into it would make it more costly than most other crashes, thus making it virtually invincible to any Profit-Based algorithm driven car. (Note, in this hypothetical scenario, the fact that one car could lie to another would eventually become public knowledge, thus making communication between cars seen as unreliable. This would either eventually become self-defeating or some verification mechanism would clamp down on this strategy.) Assuming that Car 2 can't lie to Car 1 and running into Car 2 is in fact less costly than any other viable option, then Car 2 would be in the same situation that Car 1 was when Car 1 was using the Humanist algorithm- nothing a Protectionist Car 2 could say to a Profit-Based Car 1 would alter Car 1's decision. Likewise, there would be no privacy concerns in this scenario.

**In 2017, an article that discussed how Tesla would solve a self-driving crash dilemma implies that in the case of the trolley-type crash dilemma, "it means that a Tesla car would not retake control of the wheel and swerve away from a group of people (or even brake), [even] if the driver were deliberately driving into them." Which of the three "moral judgment" algorithms is the Tesla approach closest to? Why?**

Assuming the driver is not suicidal or extremely altruistic, this falls under the Protectionist algorithm. Most drivers would instinctively try to protect themselves at all costs. However, if there is time enough for a driver to evaluate a bad situation in which someone is going to die and the driver is ethical enough to sacrifice themselves to reduce the number of deaths, then this could turn out to be a Humanist algorithm.

**What if your self-driving car must operate in multiple countries (i.e., U.S. versus Germany versus China) where driver norms can be very different from one context to another - Do**

**you think the car should adapt its "moral judgment" stay the same, or allow the human to switch from one ethical setting to another automatically? Why?**

I think in most cases, the Humanist algorithm is the most ethical; however, I could imagine going to a place where the environment is suddenly and unexpectedly very hostile wherein my life is seen as less valuable than a human's and all manner of unavoidable dangerous obstacles are thrown in my path. In this extreme case, I would want to be able to adapt its moral judgment from Humanist to Protectionist.

**Imagine that the newest self-driving startup decided its self-driving car would always follow the road rules. Based on the scenario you selected in Step 2, what traffic rule(s) (if any) did the self-driving car violate in the humanitarian, protectionist, and profit-deployed algorithms? Who are the specific individual(s) in the scenario that would be placed at harm if you followed all the traffic rules? Explain your reasoning.**

This is a legally complex question, so first I will set the stage with some basic facts and assumptions. I am not a lawyer, but it is my layperson understanding that when one car, Car A, rear ends another car, Car B, it is usually assumed that Car A was following Car B too closely and Car A is at fault; however, if Car B changes lanes or pulls out in front of Car A without giving Car A enough space to slow down, then Car B is at fault. With that in mind, Car B would be at fault if Car A (a.k.a., Car 1) rear ends Car B in the scenario I selected in Step 2. Moreover, if Car A/1 veers away to avoid Car B to instead run into a bicyclist or Car 2 knowing that doing so may kill whoever they run into, that may constitute involuntary manslaughter.

With all that in mind, it's pretty straightforward what traffic rules did the self-driving car violate in each of the algorithms. The self-driving car violated no traffic rules in the case of the Humanist algorithm; in both the Protectionist algorithm and the Profit-Based algorithm, the self-driving car would have violated traffic rules and possibly be guilty of involuntary manslaughter. Following all the traffic rules, the specific individuals placed at harm are the occupants of Car B, male age 18 and 2 males age 19, because they are the ones at fault (or at least the driver is) and thus are the ones that Car 1 is legally required to run into after making an attempt to slow down in time.