So, here's the problem with **self-driving cars**

Frank van Praat MA MSc RE
is a senior manager at KPMG.
vanpraat.frank@kpmg.nl

Almost all scenarios about the future of our mobility show that autonomous vehicles are here to stay. But the level of autonomy that will actually be achieved, also heavily depends on whether existing technical, legal and ethical challenges will be overcome. Because how much of your personal safety are you willing to trade for being on time for that one important meeting? This so-called trade-off requires making a decision between two conflicting benefits: safety and usefulness. And although it seems easy, to make such decisions work in unprecedented situations that go beyond human imagination, is a challenge that might never be solved.

INTRODUCTION

Automotive companies are making considerable progress in the development of self-driving cars. The steps taken are impressive. However: the obstacles are just as impressive. It is becoming increasingly clear that the dreamed level 5 of full autonomy might never be reached. Why not? There are multiple reasons for this, both technological and ethical. One issue really stands out: level 5 autonomous driving will require self-driving cars to continuously make trade-offs between safety and usefulness as well as making these trade-offs in interaction/communication with other vehicles and (slower) road users. Sometimes the decision will be straightforward, sometimes it will be immensely complex. The challenge is to see how a fully autonomous car will decide in these particular situations. With these types of trade-offs, it seems that technology is causing the issue on the one hand, and on the other it is obvious that the ethical and regulatory considerations are much more important. In this article we will explain why.

AUTONOMOUS DRIVING: WHERE DO WE STAND?

Big names that have focused on the development of self-driving cars: Tesla with AutoPilot, Google with the Waymo, General Motors with the Super Cruise. Also; Nissan equips its Leaf with more and more independent options, BMW launches one update of its software after another and meanwhile Ford and VW are investing billions in this technology via Argo AI. Baidu – experimenting with self-driving cabs in Beijing – has concluded cooperation agreements with at least four Chinese car manufacturers.
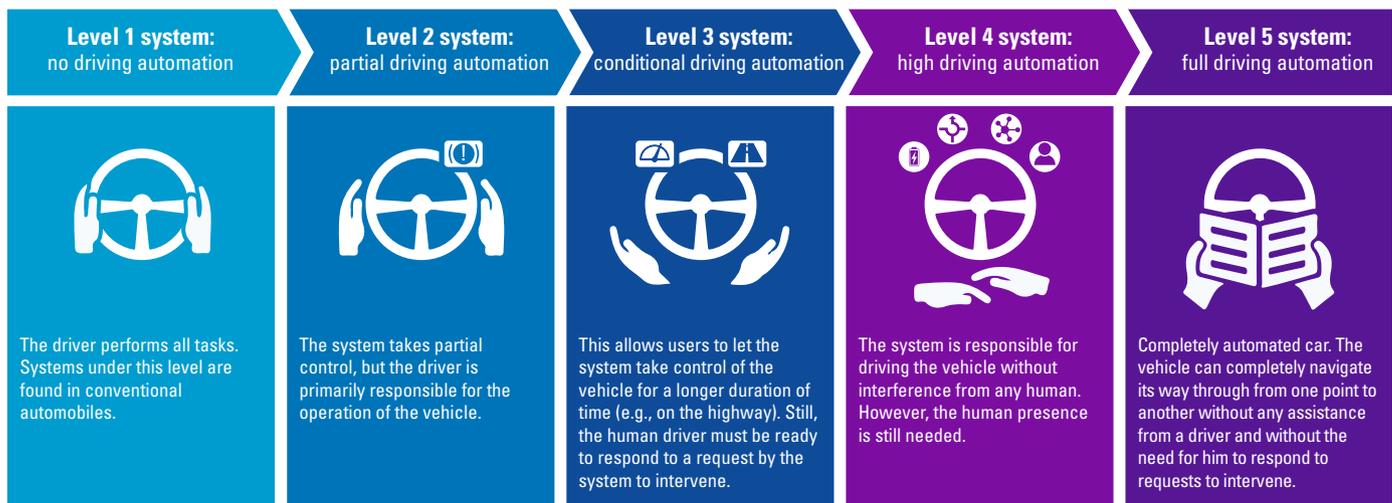
Autonomous vehicles (including cars) are here to stay. They play a prominent role in almost all scenarios regarding the future of our mobility. Sometimes they even play a leading role. In this article we examine the foundations of these expectations. Can they be met? Are we on our way to build a fully self-driving vehicle, a vehicle that functions reliably under all circumstances and meets all our mobility needs, because it is perfect in terms of infrastructure use, flexibility, efficiency and convenience?

Before answering that question, we will first put an end to the confusion in the wording regarding "self-driving vehicles". Quite a few definitions and terms are circulating. However, "self-driving" is of course a myth: cars do not move by themselves yet. Also, the terms driverless cars and autonomous vehicles are inaccurate, let alone a term like robotic cars. Then what? The Society of Automotive Engineers (SAE) examined this issue in 2018 and published a taxonomy on the subject. According to that standard, the appropriate terms are driving automation or driving automation systems. Automation is the use of electronic or mechanical devices to replace human labor. Clearly, the human labor involved is driving. Hence driving automation. (It's not the car that is automated.)

The SAE-standard is especially interesting because it de-scribes the levels of driving automation, better known as the SAE Levels. These five levels are described in Figure 1.

In 2020, we are at the second level, the level of partial driving automation. Most experts agree on this, though some manufactures such as Audi even claim to have achieved level 3; however, their claim has not yet been proven in

"Self-driving" is a myth: cars do not move by themselves yet

Figure 1. SAE Levels of driving automation.

Level 1 system: no driving automation	Level 2 system: partial driving automation	Level 3 system: conditional driving automation	Level 4 system: high driving automation	Level 5 system: full driving automation
The driver performs all tasks. Systems under this level are found in conventional automobiles.	The system takes partial control, but the driver is primarily responsible for the operation of the vehicle.	This allows users to let the system take control of the vehicle for a longer duration of time (e.g., on the highway). Still, the human driver must be ready to respond to a request by the system to intervene.	The system is responsible for driving the vehicle without interference from any human. However, the human presence is still needed.	Completely automated car. The vehicle can completely navigate its way through from one point to another without any assistance from a driver and without the need for him to respond to requests to intervene.

practice ([Edel20]). Experts also agree that the industry still has a long way to go. After all, technology is advancing, and social acceptance is also increasing. The accidents that occur involving level 1 or 2 driving automation systems are widely reported in the media, but at the same time, the applications of driving automation are in great demand. Whether it concerns cab services for short and safe routes or the next update of Tesla's AutoPilot, people are enthusiastic.

Where will this end? It is my belief that level 5 systems are an illusion. Because of one fundamental issue: deciding on the trade-off between safety and usefulness, especially for situations that initially also go beyond human imagination.

TECHNOLOGICAL TRADE-OFFS

A trade-off is a situational decision that involves diminishing or losing one quality, quantity or property of an item or product in return for gains in other aspects. In a trade-off it is impossible to have it all. In the field of emerging technologies, trade-offs are very common. Think of large enterprises such as Google that provide you with all kinds of services like Google Maps and Google Drive "for free". Obviously, there is no such thing as a free lunch. This means that for all the so-called free services, you have to make a trade-off and share your data in return. Another example of a trade-off in technology concerns our mobile phone. On the one hand, it gives us freedom, and we seem to not be able to live without it. On the other hand, we've become so attached to it that we lose the ability to get the same benefits out of other things in our lives.

These examples show that technological trade-offs come in all shapes and sizes. From relatively high level and almost philosophical, to very concrete and technology specific. Automated driving systems, for example, are famous for using Artificial Intelligence (AI) and advanced algorithms at scale. But when these algorithms are developed, it is also very common to make design decisions that require making a trade-off. Let's have a more detailed look with some examples.

The *performance* versus the *explainability* trade-off

One of the most well-known algorithm trade-offs is between performance and explainability. Performance in simple terms is the extent to which a predicted outcome of an algorithm matches the actual situation: the better the match, the better the performance. Explainability means the extent to which an input-output relationship of an algorithm's outcome can be explained. Easier said: on what grounds was a certain outcome reached? Based

on practical experience as well as scientific literature ([Vilo20]), we have learned that it can be a real challenge to optimize both performance and explainability for a single algorithm. Because the more data you put into a single algorithm, the more variables it will be able to weigh, which most likely will result in better predictions compared to one using less data and variables. Using so much data and variables as input comes with a price, however, as at the same time, the different variables will become much more difficult to oversee and require exponentially more calculations. This aggravates the task of explaining how they were exactly weighed in case of a certain algorithm outcome. Especially when the number of variables goes beyond human imagination. Something that is not uncommon in algorithm development.

The *individual* versus *group* trade-off

This algorithm trade-off is best explained with an example. Consider yourself driving on a very busy and almost congested highway. Your intelligent navigation system predicts that the busyness ahead will most likely cause a traffic jam that will cause at least a 30-minute delay. Therefore, the system pro-actively offers you an alternative route that is much faster, using multiple smaller and less busy roads through a variety of quiet villages. However, you are not the only one driving on this busy highway in a car with such an intelligent GPS. Imagine what will happen: every driver using this system is offered the identical detour, because at an individual level it looks like the best option. Yet in the meantime, the impact on the group will be very negative, because the alternative route will become very crowded quickly, potentially resulting in even more delay. Let alone the nuisance for the villages' citizens, caused by the detouring cars and heavy trucks.

The *performance* versus *fairness* trade-off

The final example of algorithm trade-offs we will discuss is the one between performance and fairness. We have explained performance in the context of algorithms a bit earlier: the extent to which a predicted outcome of an algorithm matches the actual situation. The term fairness is also very common in the field of artificial intelligence. In short, it means the extent to which a predicted outcome is fair in similar situations for similar (in most cases protected) subgroups. But the decision how to measure fairness is difficult and highly context-specific, e.g. the impact of false positives[1] in credit risk scoring or criminal recidivism prediction versus the impact of false negatives in medical diagnosis. In general, fairness hurts perfor-

1 A false positive score in AI means a that there is a positive discrepancy between the predicted outcome (for example 'yes') and the actual outcome (for example 'no').

mance because it diverts the objective of an algorithm from performance only, to both performance and fairness combined.

The *safety* versus *usefulness* trade-off for driving automation systems

As mentioned before, algorithms play an important role in driving automation systems. It therefore goes without saying that trade-off decisions are most likely to be present here as well. But before we discuss this further, we must first understand how driving automation systems actually work.

Basically, algorithms are used in these systems for two main functions ([KPMG16]). Firstly, there is *observation*. Driving automation systems collect all sorts of data from radars, cameras, lidars and ultrasonic sensors. Radar technology is used to monitor the position of nearby vehicles. Video cameras help to detect traffic lights, read road signs, track other vehicles, and look for pedestrians. Lidar (light detection and ranging) sensors bounce pulses of light off the car's surroundings to measure distances, detect road edges, and identify lane markings, and ultrasonic sensors in the wheels detect curbs and other vehicles when parking. All this data is combined into a representation of dynamic objects and static features around the vehicle. Via advanced algorithms, detailed pre-generated maps are

Example 1

A car with a level 5 driving automation system approaches a pedestrian crossing. There is a woman on the sidewalk, she is standing still and making a phone call. At first glance, it doesn't look like she is preparing to cross the street, but you are not sure. What should the system do? Obviously, the safest option is to slow down, stop, and wait for a few seconds to give the woman the opportunity to cross the street. If she does, the decision turns out to be the right one. If she does not cross the street, the system made the car stop for nothing.
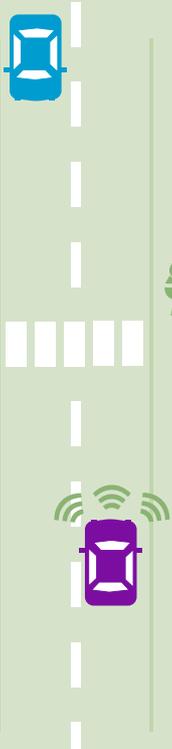
This is what the safety versus usability trade-off comprises. The safest option is to stop, even when it turns out to be for nothing. The user-friendly option would be to continue driving, with a solid chance that it would result in an accident.

In real life, humans are able to carefully read the situation and make a balanced decision based on their intuition. Also, we have the possibility to make eye contact, which can help us to decide what to do. But a driving automation system does not have these capabilities and possibly never will. Therefore, trade-off decisions between safety and usability will be needed all the time. And safety will in most cases be favored over usability. But against what cost? Would you buy a car with level 5 driving automation that keeps stopping for nothing?

Example 2

A car with a level 5 driving automation system approaches a crossing and wants to turn right on a road with an 80 kilometers per hour speed limit. When the car arrives at the crossing, the system makes it stop and uses the car's cameras, radars and sensors to determine if it is safe to turn right and enter the 80-road. There still is an uncertainty, however: the technological assets of the car can only scan for a certain distance and will possibly miss approaching cars that are exceeding the speed limit. In order to deal with this uncertainty, the safest way would be to enter the road by accelerating to 80 km/h as quickly as possible ([DeTe19]). But although this might be safe, it would be a very surprising and probably not very comfortable event for the person inside the car. The sense of usability is again under pressure, mainly because the car feels very unpredictable. Once again, we need to decide on a trade-off between safety and usability.
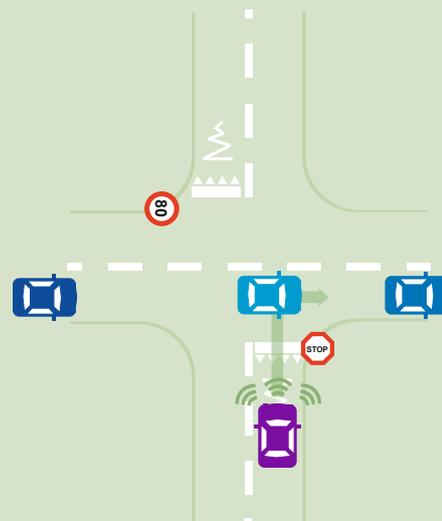
enriched to precisely compute the current position of a vehicle, representing position and orientation.

Secondly, there is *control*. Based on the surroundings and the position of the car, the systems in the vehicle need to "decide" on the best next action. This decision is based on a combination of hard-coded rules, obstacle avoidance algorithms, predictive algorithms and object recognition algorithms. Collaboratively, they send instructions to the car's actuators, which subsequently control acceleration, braking, and steering.

Based on the observed situation, a driving automation system should be able to decide what to do, which is a continuous process as the system's observations also continuously change. But when making all these decisions, there is also a trade-off to make which we introduced in the beginning of this article: the decision between usability and safety.

SO, HERE'S THE PROBLEM WITH "SELF-DRIVING CARS": THEY WILL NEVER BE BUILT

Cars with driving automation systems are required to decide on these trade-offs almost continuously. Each fraction of a second a new situation is observed that needs to be controlled. The good thing is that driving automation systems are exactly built to do so, and in most situations, there will be no issue at all. But in unprecedented situations that never happened before, the control systems are really put to the test. Should the system favor safety over usability, or the other way around? To decide on this compromise there needs to be some kind of framework that will help the system to make a humanly acceptable decision. What that framework should look like is something that needs to be decided by society as a whole. But the regulatory and ethical considerations will be tremendous. This is why the question arises if we will ever be able to create such a framework.

From my perspective it will be almost impossible. Because of two main reasons:
- The framework should be the result of an extensive moral discussion and help driving automation systems to make decisions that are acceptable for all humans. But as the moral landscape stretches out over so many values such as freedom, human autonomy, human equivalence, pleasure, safety, health, etc., there will always be a collision of these values that can only be solved via democratic systems. We believe that making these democratic systems work at the required, potentially global, scale will never happen.
- The framework should also account for situations that have never happened before and most likely go beyond any human imagination. Therefore, the framework will never be done and be complete.

Is that disappointing? No. After all, the developments in driving automation are promising enough and there is nothing wrong with a thoughtful combination of man and machine. Hybrid Intelligence seems to be the solution here, using technology to augment human intellect. Not replacing it but expanding it.

In my vision this will result in driving automation systems that:
- can be fully autonomous in situations with proven limited trade-offs (e.g. driving on the highway);
- will timely give control to humans in situations where typically too many trade-offs are expected[2] and human intelligence is required to take over.

2 Situations that self-driving car cannot handle autonomously are called edge cases. There are several possibilities to overcome these edge cases, having a dedicated infrastructure for the autonomous car is one of the alternative solutions.

References
[DeTe19] De Technoloog (2019, December 5). De autonoom rijdende mens. BNR podcast. Retrieved from: https://www.bnr.nl/podcast/de-technoloog/10396604/de-autonoom-rijdende-mens

[Edel20] Edelstein, S. (2020, April 28). Audi gives up on Level 3 autonomous driver-assist system in A8. Retrieved from: https://www.motorauthority.com/news/1127984_audi-gives-up-on-level-3-autonomous-driver-assist-system-in-a8

[KPMG16] KPMG (2016). *I see. I think. I drive. (I learn). How Deep Learning is revolutionizing the way we interact with our cars.* Retrieved from: https://assets.kpmg/content/dam/kpmg/tr/pdf/2017/01/kpmg-isee-ithink-idrive-ilearn.pdf

[Vilo20] Vilone, G. & Longo, L. (2020). Explainable Artificial Intelligence: a Systematic Review. Cornell University. Retrieved from: https://arxiv.org/abs/2006.00093

About the author
Frank van Praat MA MSc RE is a senior manager at KPMG and leads KPMG's Trusted Analytics team. He has extensive knowledge about Emerging Technology Risk Management, Advanced analytics governance and auditing, and the human side of AI.