Utesch et al. European Transport Research Review (2020) 12:48
https://doi.org/10.1186/s12544-020-00438-2

European Transport Research Review

REVIEW Open Access

Towards behaviour based testing to understand the black box of autonomous cars

Fabian Utesch*, Alexander Brandies, Paulin Pekezou Fouopi and Caroline Schießl

Abstract

Background: Autonomous cars could make traffic safer, more convenient, efficient and sustainable. They promise the convenience of a personal taxi, without the need for a human driver. Artificial intelligence would operate the vehicle instead. Especially deep neural networks (DNNs) offer a way towards this vision due to their exceptional performance particularly in perception. DNNs excel in identifying objects in sensor data which is essential for autonomous driving. These networks build their decision logic through training instead of explicit programming. A drawback of this technology is that the source code cannot be reviewed to assess the safety of a system. This leads to a situation where currently used methods for regulatory approval do not work to validate a promising new piece of technology.

Objective: In this paper four approaches are highlighted that might help understanding black box technical systems for autonomous cars by focusing on its behaviour instead. The method of experimental psychology is proposed to model the inner workings of DNNs by observing its behaviour in specific situations. It is argued that penetration testing can be applied to identify weaknesses of the system. Both can be applied to improve autonomous driving systems. The shadowing method reveals behaviour in a naturalistic setting while ensuring safety. It can be seen as a theoretical driving exam. The supervised driving method can be utilised to decide if the technology is safe enough. It has potential to be developed into a practical driving exam.

Keywords: Deep neural networks, Autonomous cars, Validation

1 Introduction

At the moment research and development of autonomous cars is popular within the automotive research domain. It is possible that autonomous driving systems (ADS) lead to a safer, more efficient, sustainable and convenient traffic environment. It is assumed that driving becomes safer by avoiding crashes caused by human error like driver impairment through distraction or fatigue. Additionally travelling could become more efficient and sustainable by offering easy car sharing and avoiding traffic jams. This could be achieved by means of better communication between ADS, other road users and traffic infrastructure. Driving might also become more convenient by relieving the driver of responsibility and reducing effort by transitioning the role of a driver to a passenger.

However, the road towards this future will not be easy. On the one hand it is possible that ADS will reduce traffic congestion due to a decrease of individually owned vehicles and coordinated driving in platoons [34]. On the other hand congestion could also increase in locations where driving is cheaper than parking [34]. Even though humans are not the perfect drivers they are still the best around. They are managing most situations without crashes. ADS will have to catch up first. This will take time and there are several technical challenges that have to be addressed first. For example, while it

* Correspondence: fabian.utesch@dlr.de
Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR) - Institut für Verkehrssystemtechnik, Lilienthalplatz 7, 38108 Braunschweig, Germany

seems trivial for humans to recognise objects, it is a fairly challenging task for computers. The main reason is that object detection algorithms are based on appearances which can be influenced by factors such as weather, scene illumination, sensor noise, occlusion and object classes [20]. Driving a vehicle heavily depends on visual information. It is possible to identify specific information about the environment by other means. For example radar outclasses camera and eye vision of vehicle detection in rainy weather. However, it has less semantic information compared to camera images [5]. Laser scanners are well suited to measure the distance of obstacles around a vehicle with high precision but lack the ability to recognise semantic information like traffic signs, traffic light status or intentions signalled by other road users. Moreover, laser scanners remain expensive compared to cameras and radars [5]. While driving, a broad range of information can be collected by the range of the electromagnetic spectrum which is visible to the human eye. In fact, streets are designed to support navigation within this visible range (e.g. traffic signs, lights and street markings) of light because they are meant to be perceived by the human eye. This is an advantage for vision based systems and humans rely almost solely on this channel to navigate through their environment, including driving a vehicle.

Visual information is extremely complex and thus difficult to interpret. Still humans manage to identify a vast range of objects, predict their movements, recognise and interpret all different kinds of signs, even if these are dirty, damaged or occluded. They identify driving related objects, potential hazards and non-verbal messages from other road users about their state and intention. Transferring these abilities to a computer has great potential to collect the information needed by cars to drive fully autonomous. Deep neural networks (DNNs) play an important role in achieving this vision. However they have the disadvantage that the decision logic is hidden from external review. This poses a new challenge: Assessing the safety of this technology. If a computer should be allowed driving a vehicle it needs to be safe. If the code does not reveal the decision process, its abilities to handle traffic situations safely need to be demonstrated in another way. Finding a satisfactory alternative is a major challenge in receiving regulatory approval for ADS. There are many approaches, but there is a gap in predicting the ability of ADS to handle driving related tasks outside of known situations. To address this, it is necessary to show that ADS can react to the semantic class of a situation independent from individual involved objects and actors. We argue that it is possible to generate behaviour based models describing the internal processes on a high level, revealing said abilities of ADS. In this paper we propose the use of four alternatives to the source code analysis to describe, understand and predict the behaviour of DNNs. The presented ideas differ from many existing approaches in that they are solely based on behavioural experiments. They should be seen as proposal for better understanding the actions of DNNs and predicting the scope of its abilities. The presented approach can be used as starting point to evaluate the presented methods. Describing a complete behaviour based regulatory approval process of DNNs is out of scope of this work. Nonetheless, the regulatory approval of ADS will be used as an example of putting the knowledge of behavioural modelling of DNNs to good use.

In the following, a brief introduction into DNNs and their relevance for ADS is given. Then the idea of behaviour based testing to approach understanding black box ADS is outlined. Four approaches to understand black box technical systems for ADS are described next. Some disadvantages of focusing on driven kilometres to measure the performance of ADS are presented. Finally, a possible way to apply these four approaches is explained.

2 DNNs as key factor for autonomous driving

Recent advances in the application of DNNs make them a promising technology for achieving ADS. Especially computer vision is a task in which DNNs are superior to classical information processing approaches. Their structure is fundamentally different to other computer software and more similar to a human brain. While classical software defines all possible decisions and actions in source code, the code of neural networks mimics a brain by defining virtual neurons that are connected in a net structure. After the code is written, decision processes and actions are generated by adding weights to the connections through training of the network. Instead of writing code that describes how the system works, DNNs are written to learn by themselves by means of training. The final challenge in safe ADS is to react in the right way to any situation at any time [45]. DNNs seem to be the best option to address this challenge.

Outside the automotive domain it was demonstrated that DNNs can perform better than humans, for example in lip reading [2] or games [11]. This is possible because DNNs are able to take many aspects into account, even some that cannot be perceived by humans. World class competitive gamers report that DNNs have developed strategies that are superior to everything that humans have thought of in their domain [27]. Some do actually report that they learn from DNNs to improve their skills in the field [12]. Looking back, computers have surpassed human capabilities in challenges of increasing complexity like Chess [28], recognition of handwriting [29], Jeopardy [19], Go [11, 12] and recently StarCraft II [54]. Considering the pace in which this evolution took place, it seems possible that DNNs might be

able to even master driving in the real world. DNNs seem well suited to do this since with proper training, they develop their own strategies to solve problems. This allows them to surpass the abilities of their creators.

The achievements of DNNs in other domains are promising. However, driving on roads is much more complex than playing games and the consequences are more severe if something goes wrong [31, 55]. While the playing field is well defined in Chess, Jeopardy or Go, each situation on the road is truly unique. Additionally, any error can cost lives. This puts high demands on the accuracy of ADS. To ensure that these vehicles are safe approaches are needed that allow assessing the capabilities and limits of ADS.

The source code of DNNs does not indicate how the system would react in a given situation; instead it describes how the system learns to function in the desired manner. The decision logic is built by letting the system learn from training material. Even though the meaning of the training set is given by humans, the interpretation of the given categories is built by the system. The logic itself is deterministic in that it produces the same decision under identical circumstances. However, the sensory input of driving in the real world is usually noisy and very complex so that two situations that look exactly the same to a human might look very different to a machine. Thus the system might react differently in seemingly similar situations giving the impression of not being deterministic. Since these decisions are expressed implicitly in the connections of the neural net and not explicitly in the source code, it is not possible to rule out errors of the system like it can be done in other software. There are efforts to get glimpses of the decision logic. For example there are attempts to identify corner cases for DNNs by generating test inputs [41, 52], visualizing what influenced a classification [7, 9, 56] and to asses if a prediction can be trusted [43]. However these approaches do only extend the list of specific cases that the reaction to is known or show what influenced a decision. They do not predict the ability of DNNs to handle certain types of situations, which is what is needed to allow ADS to drive in real traffic.

3 Relevance of behaviour based testing

To receive regulatory approval it is necessary to show that ADS are safe. The first step in this direction is their validation. There are already efforts to develop tools that measure the performance of ADS and quantify it in comparison with other systems. For example, project PEGASUS, funded by the German Bundesministerium für Wirtschaft und Energie, had the specific goal to develop tools to accelerate the regulatory approval of ADS for the mass market [14]. It focuses on the approaches virtual testing, proving ground tests, field tests and decomposition [49]. Each method contributes its strengths to generate a holistic view of the issue. The presented approach here differs in that it focuses exclusively on ADS that use DNNs, it strives to create a schematic model of the function of these DNNs and it focuses exclusively on working systems.

The analysis of not directly observable whole systems, functions or elements of systems is part of the psychologist expertise since a long time. Many psychologists during the fifties were focused on behaviourism. It specialises on analysing, understanding and explaining observable human and animal behaviour by objective methods of science while assuming the brain is a black box. Underlying cognitive or physiological mechanisms were out of scope. Later on cognitivist psychologists developed behavioural experiments to analyse exactly these underlying mechanisms and therefore to measure the performance of the mind. Some of these methods may be suited to evaluate DNNs of ADS.

There is indeed already a behaviour based solution to evaluate the ability to safely drive a car: The driving test for humans. The brain is basically a black box where the decision logic is mostly unknown to the outside. It is arguably unknown why drivers drive as they do and which decision rules they follow. There are approaches to explain cognitive processes but these are models for human behaviour without actually seeing the algorithms of the brain itself. The models are an approach to extract these algorithms by approximation, which is the analogy that is proposed here for ADS. The theoretical and practical driving tests were developed to evaluate the driving skills of new drivers despite the fact that the human mind is hidden from external review. Due to the differences between a human mind and a DNN a direct appliance of the human driving test to ADS is unlikely to yield the same effect. However, when adapted accordingly, it could turn out to be a useful method for assessing the functionality and safety of ADS. In the following, steps to transfer the human driving test to the context of ADS are outlined.

4 Alternatives to source code analysis

Static source code analysis is a common way to evaluate the safety or correct function of software. This approach does not work for DNNs because its source code does not describe its input-output behaviour, but its learning of making its own decisions which are not part of the original programming. When reviewing the source code is not an available option, there are at least four different approaches that do not require a direct view on the decision logic. Each one is not exclusive, but instead should be seen as complementing the others.

The experimental psychology method uses behavioural experiments to create pragmatic models of the internal

decision logic of the ADS to understand its function in regard to traffic safety. Penetration testing utilises available information about the ADS to provoke errors with as little effort as possible. This knowledge can be used to understand the limits of ADS. It can also be utilised to improve the system and make it robust against such attacks. In the shadowing method the ADS is built into cars, but not activated. The systems performance in everyday situations can be observed without allowing any real action. Its decisions can be compared to the performance of the human driver to understand its capabilities. Finally in supervised driving ADS are driving by themselves in everyday traffic. They have to show that they can react safely in any situation under natural conditions. They are kept under supervision as failsafe against errors.

These methods can be used to understand and improve ADS (if they fail) or strengthen trust in their safety (if they pass). Each method is presented in more detail below.

When methods of experimental psychology are to be applied on DNNs it has to be kept in mind that there are many differences between human drivers and DNN based ADS. Thus an adaptation of the testing procedure makes sense. Both machines and humans do have key advantages in regard to driving vehicles. Humans can generalise. Humans only require around 30 h of driver training to master the driving exam. However, at the point of the exam humans already have about 16–21 years of experience navigating within the world and 10–15 years of navigating in traffic as a road user. Even if the previous experience is based on walking as a pedestrian or riding a bike, humans are capable of generalising the learned rules to the new domain, like driving a car. Traffic lights still regulate behaviour, lane markings still mark the areas one is allowed to move on, signs still indicate what rules have to be followed. Usually a few key situations are enough for humans to be able to generalize from. Machines have to be trained with vast sets of right-wrong examples to derive a decision rule that is still specific. Humans learn by observation. Humans are able to learn by observation of others solving a problem. Machines are only capable of this in a very limited way. Humans learn by instruction. Humans can be given the decision rules directly (driving school), instead of inferring them by trial and error. Technically the same is true for machines. However the instructions have to be much more specific for machines which limit the applications.

On the other hand machines are fast learners. Due to higher frequencies of processors compared to firing neurons and higher signal speed compared to nerves, machines are able to speed up the simulation of situations by several orders of magnitude. This allows them to find the best way to handle a situation by trial and error in time frames that are impossible for a human to match. Machines can learn in parallel. It is possible to multiply the code of a system to parallelise the learning process. This enables machines to learn even faster or react to a copy of itself to simulate interaction with another entity. Machines do not get tired. While humans need to rest to replenish their strength, machines are able to keep learning in simulations for days, months or years, as long as power and maintenance is provided. Machines can copy knowledge. Useful abilities can be copied to new machines. Also, DNNs can be stacked and an expert system can be used as input for a higher level DNN decision system.

Self-driving cars do not have the same experience or abilities as humans. They usually start from scratch and have to catch up with all that a human has learned already. However, machines do learn certain tasks much faster than humans, can access a broader sensor battery that is at least in parts superior to human senses and utilise information from more sources than just five senses. All these differences underline that the driver test method has to be adapted to be utilised for ADS.

4.1 Understanding the system

At the moment ADS are still new and unfamiliar. How can trust in such a system be gained when the known methods are not suitable for testing it? The short answer is to build it up over time. Long hours of observation of real traffic can help to reach a level where the system is able to drive in many different traffic situations. The first step is to understand the abilities and limits of ADS.

4.1.1 Method of experimental psychology

To better understand DNNs and their behaviour a more systematic approach can identify limits of ADS by modelling its internal function. In this method the function of ADS is modelled in experiments solely based on the reaction to different stimuli.

Contrary to robots, human minds do not have a known source code and their decision rules are not directly observable. The brain is like a black box, its logic is hidden to a human observer. There are imaging techniques like electroencephalography (EEG), regular and functional magnetic resonance imaging (MRI, fMRI) or positron emission tomography (PET) to observe the brain. However these provide either a high temporal or spatial resolution, but none of them both. Unfortunately these neuroimaging techniques are not yet detailed enough to understand decision processes. To comprehend, evaluate and predict human behaviour, psychologists developed behaviour based methods instead. Experimental psychology strives to describe human thinking by means of observation of the reaction to

specific conditions. This is achieved by creating very specific tasks to learn about a function. The first step of the model building process is the observation of behaviour of the mind; an assumption is made and then tested. The initial model is then extended by testing more assumptions that are derived from what is already known. Every step is documented, so that the procedure can be replicated by anyone to verify the validity of the model. This process is demonstrated in a simplified example about understanding human memory.

Humans are capable of remembering information. If asked, subjects are able to recall the information and their origin from the last month. Apparently, there is something inside the brain that is capable of retaining information: memory. But humans also forget information over time. This can be visualised in a very simple model (Fig. 1).

Experimental psychologists conducted many experiments to investigate memory, its capabilities and limits. By learning nonsense syllables and documenting the forgetting curve, Ebbinghaus found that most information is forgotten immediately after learning. Information that survived the first week can be retained for a long time [16, 37].

To find out how much information can be stored, a row of numbers was shown briefly and subjects were then asked to recall it. People were able to recall between 5 to 7 words immediately [35]. The model of short term memory can be enhanced accordingly.

The size of the short term memory indicates that a row of numbers like 6 8 4 9 5 1 3 3 7 5 4 5 will almost never be remembered fully without extensive learning. However additional studies using different modes of presentation revealed that altering the row to 684,951, 337,545 makes it much easier to remember completely [10]. It is the case even though they are the same numbers, only grouped. This indicates that the same information can be stored more efficiently if the meaning is changed [35]. This process is called 'chunking' (Fig. 2).

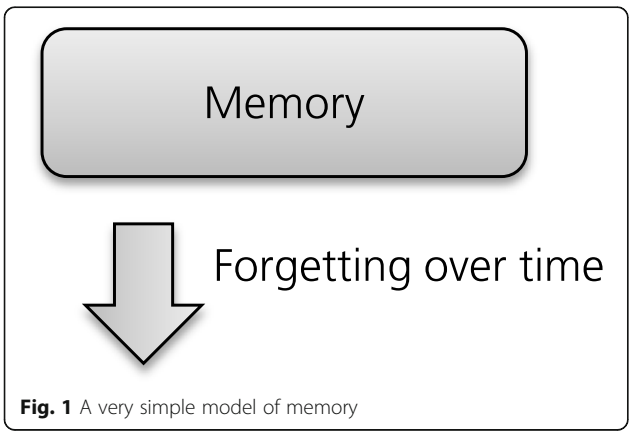

Fig. 1 A very simple model of memory

By performing more studies changing the type, position, occlusion, duration and presentation of stimuli psychologists were able to create a sophisticated model of human memory (Fig. 3). This was done without ever being able to see the decision logic inside the brain. All that was needed was a neat selection of the stimuli presented and the exact documentation of the reaction during recall [3, 4]. Actually much more is known about memory, but this simplified model shall suffice to demonstrate the process of behaviour based learning about hidden functions.

Even though it is unclear exactly how human memory works, scientist were able to describe its function by means of size, type and duration only by observing the reaction in certain tasks. The resulting schematic model is sufficient to make predictions about the abilities and limits of human memory. Even shortcomings can be predicted. The better the systems actions can be understood, the better its actions can be predicted in hazardous situations.

To give an example: A person will probably have a hard time remembering a password of 12 random letters, numbers and signs. But the same person should be able to easily remember a password of 5 words like pony – tantrum – headset – spring – spaceship [36]. Even though the first password is shorter (12 characters) than the second one (33 characters), the latter should be easier to remember, according to what is known about memory. Incidentally, this would also increase the security of the password since length adds more entropy than increasing the vocabulary of allowed characters. It is not necessary to look into the black box of the brain to draw these kinds of conclusions. The same method could be utilised to improve ADS or assess its safety.

The experimental psychology method can be adapted to understand the function of DNNs. Like a human, ADS can be presented with different stimuli to measure its reaction. By clever choice of input stimuli the internal processes of DNNs can be derived step-by-step and compiled into a model. That is to say not the actual internal structure, but a high-level representation of relevant processes.

For example investigating the processing of visual information could be realised through presenting driving related objects (traffic signs, lane markings, other road users and obstacles) with different levels of occlusion and measuring the reaction of the system compared to the expected behaviour (adapting speed, steering, swerving or braking). This would allow investigating the information acquisition as well as the information processing. It would be a viable option to assess the safety relevant capabilities of a system.

To make statements about the safety of ADS it is necessary to predict its behaviour in a wide range of

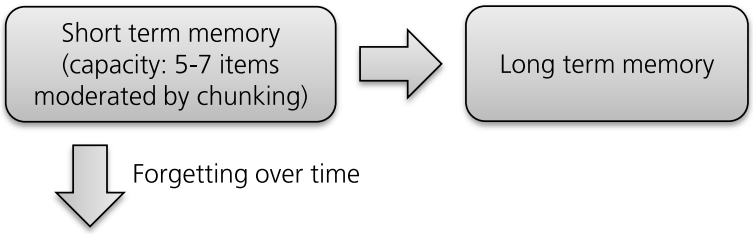

Fig. 2 Model of memory with forgetting and chunking

situations. Since it is impossible to test DNNs in every possible situation, the logic of the decision needs to be understood sufficiently well to extrapolate the behaviour from a selection of situations. The reactions allow deriving abilities and limits of DNNs. They can then be improved to avoid these mistakes in the future. The method of the experimental psychology was used to create a model of the human mind. It can similarly be used to model the decision logic of DNNs. However, knowledge is gathered in small steps. It will take time to design and perform the experimental studies. Utilising the advantages of machines over humans may accelerate certain steps, but this process has to be developed first.

Additionally ADS are improved continuously. Not only may ADS gain new abilities and handle situations better, it is also possible that previously achieved abilities are removed by a system update. This makes it difficult to assess since they may change with every update. Due to this fact an assessment for ADS can only be made for a specific software version. While it may not appear feasible at first, the development of paradigms to investigate specific ADS abilities could decrease time and costs of applying this method. This would also make it easier to retest newer versions of the software with reduced effort.

4.1.2 Penetration testing

The safety of ADS can also be tested by trying to provoke unsafe behaviour and use this knowledge to improve the system. It has to be assumed that all kinds of events could happen in real traffic. Thus it is important to bring ADS to their limit even before bringing them on the road. Penetration testing is already a common practice in computer science. It is usually applied to any kind of technology by individuals or research groups to draw attention towards security or safety risks. The idea is to think of an easy way to provoke a dangerous reaction within a safe and controlled environment. Such "attacks" can be directed to any point of the system. For example a penetration tester could try to confuse the vision system of the vehicle by making changes to traffic signs [18] or lane markings.

For example to test ADS reaction under the condition of poor visibility, dirt, glare, occlusion, rain or wetness can be reproduced relatively easy with tape, water, mud or a flashlight. All these manipulations limit the sight or distort the image. Weak spots can be identified by systematically varying the position and extensiveness of the occlusion. The reaction reveals computer vision as well as information processing capabilities.

Another attack vector would be to send false messages to ADS over car to x (C2X) channels or alter the reflection of radar on vehicles. In any case they must be able to cope with the situation, at least by safely disengaging from the traffic. The better ADS cope with the attacks, the better they will handle themselves in other edge cases. This strategy is also effective in hardening the vehicle against people with malicious intent. An American car manufacturer has

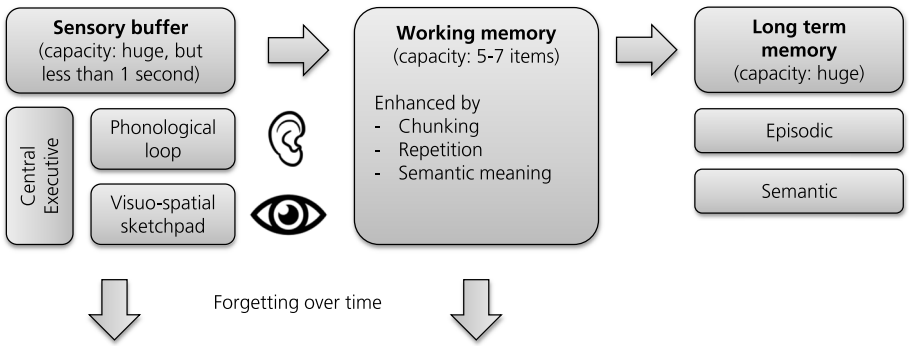

Fig. 3 A more detailed model of human memory

recently announced to actively support this kind of safety and security research [15, 50].

Even though testing ADS under real conditions is preferable because of the high validity, it is also cumbersome and slow. A very fast alternative is simulating the systems input inside a computer. The development of simulation based approaches that test behaviour of DNNs in edge cases already exists. Pei et al. [41] have developed a method to specifically target neurons of DNNs by their activation function to find stimulus combinations that trigger an unusual behaviour by the DNN and there are also efforts to specifically simulate adverse conditions like fog or rain [52]. Another approach between simulation and reality is mixed-reality in which virtual objects are injected into real world data perceived by the ADS [13]. Such methods can help in identifying weak spots of the system. If only the reaction towards a situation is to be investigated a simulation based on known positions of other road users would suffice [32]. If the object recognition itself should be investigated, a photorealistic graphical input has to be provided towards the system.

Aside from investigating the conditions around the vehicle, the functions inside of the vehicle can be investigated as well. Components can be deactivated or disconnected systematically to better understand the systems functions inside the vehicle modules. This is important since the system also has to react safely when some of its components are breaking down or are damaged. Every weakness that is understood in this way can be addressed to strengthen ADS.

Many companies do not appreciate such practices since it can be viewed as exposing flaws in their products and puts pressure on them. However, some companies see this input as valuable contribution. These companies explicitly hold hackathons in which volunteers compete in a formal setting to find critical vulnerabilities in their products [1]. The winners are formally announced and either financially compensated or receive the product as trophy. This approach is well suited for improving ADS since it reliably identifies vulnerabilities in a trusted environment. It has a high objectivity and validity since the method is published after the vulnerability has been fixed. It also has a high practicality since the attack vectors are the same as if the product would be out in the field. It can be applied to any ADS. Independently of the official adaptation of this method it is likely that independent groups will perform penetration testing on ADS as soon as they are released. Thus it is even more important that this method is applied in a controlled environment as internal testing usually done by subcontractors, institutional testing by research facilities or public testing in form of the described contests.

4.2 Building confidence in ADS

After thorough testing ADS have to go back to development to be improved. The method of experimental psychology and penetration testing can actually be seen as a loop that needs to be iterated until ADS are assumed to handle real traffic situations. When ADS are understood well and assumed capable of driving on their own, it is time to test their performance in actual traffic.

4.2.1 The shadowing method

This approach describes how ADS can be evaluated in everyday life without jeopardising the safety of other road users. The gained knowledge is utilised to train DNNs for autonomous driving. ADS are implemented in every aspect. But instead of letting them drive the vehicle, the decisions of ADS are recorded. The record is then compared to the actions of the human driver.

An example would be that a human driver exceeds the speed limit and drives too close to another vehicle. Then the other vehicle brakes and it nearly comes to a crash. An unusual abrupt deceleration and proximity warning could trigger the investigation of this incident from the collected data. It is likely that the preceding high speed and short distance to the other vehicle will be identified as a contributing factor of the safety critical event. The system was observing this situation and indicated for the record that it would have driven at a lower speed with a safe distance. While we cannot know if the safety critical event would have been avoided by this action, this would indicate that the system made the safer decision, compared to the human driver.

The kilometres driven in shadow mode will be an indicator about the abilities of the system. A possible milestone for a high confidence would be when the shadowing data shows that the system makes significantly less errors than an average human driver. If ADS decision were different to the human driver and dangerous, this would indicate that it still needs to be improved. If the decisions are safe and similar to the human over a long time and in many vehicles, the trust in the system would grow.

This method is able to identify ADS behaviour in all kinds of unknown situations without endangering anyone. It can be applied in parallel on many vehicles to speed up the process of training ADS. It is most effective if applied in the setting of naturalistic driving studies [6, 17]. Here, common human drivers are considered in everyday situations. Only when drivers are using ADS in their everyday lives it is possible that no relevant situations are excluded. It has the advantage to keep the training data as natural as possible, which grants a high validity. This is actually done by original equipment manufacturers (OEM) [21, 44]. Of particular interest are the situations that are difficult for ADS to solve. They

are rare and thus challenging to train on the street. Thus it may be useful to stage these situations under controlled conditions to improve ADS [39]. The sample should be as large as possible, because the quality of the trained DNN improves a lot the more data is used for the training and safety critical events are scarce.

4.2.2 Supervised driving

In a driving test the capability to drive a vehicle safely is measured and then based on the performance a driving license is granted or denied. The basic idea of the supervised driving method is that a human driving test can be adapted to be suitable for ADS. An original driving test for humans in driving school (in Germany) usually looks like this: After thorough training in driving related tasks a driving test is taken. In it, the applicant has to drive a vehicle safely under everyday conditions and the scrutiny of the examiner. The applicant has to master every emerging situation to pass the test. Since the exact testing conditions are unknown, the applicant has to prepare for all potential situations. An analogy could be supervised driving of ADS. In supervised driving ADS drive by themselves, but have a human as failsafe to take over when needed.

Because machines are different from humans in many aspects, the driving exam has to be adapted to cater the specific needs of ADS. Humans are only allowed to have a driving license if they have reached a minimum age, are capable of driving a vehicle, were properly trained and passed theoretical and practical driving exams [33].

How much supervised driving is necessary to prove that ADS are safe? It is difficult to set a specific limit to indicate the threshold of a machine driving better than a human driver. A popular metric to compare ADS and human driving are kilometres driven without intervention, crash or death [33]. Maurer et al. [33] estimate that approximately 6.62 billion kilometres need to be driven to sufficiently prove the safety of ADS. Usually the process has to be repeated if the system is altered. However there can be exceptions when the new system varies only in minor aspects from the homologated system. If these estimates are only somewhat true, it will take a lot of resources to evaluate ADS. Several companies seem to have understood the necessity of supervised driving to achieve regulatory approval of ADS and are using it already like Tesla [46], Waymo [30, 32] or Uber [24].

4.3 The value in testing abilities

The kilometres driven under supervised driving are not all equal. The distance driven without incident can only be representative for the situations that were covered during the training. This means that distances driven in an artificial or selected scenario are less valuable than distances driven in unfiltered natural traffic. On the other hand this means that any supervised driving that is restricted within a certain private property, city or country can only be assumed valid or safe for this scenario and not be generalised towards other areas. So it is not just important to drive enormous distances without incident, but also all potential areas, that the vehicle should drive in later, need to be covered. If ADS are supposed to drive to every place this would mean it has to be tested in natural traffic.

Driven kilometres also do not reflect encountered risk. They could be complemented by reporting the covered situations (e.g. intersections, roads shared by cyclists and vehicles). A distance that was driven in known high risk areas says more about the abilities of ADS than the distance driven on safe roads. A more extreme approach is to put ADS in the most challenging situations to see if they can handle them. This was the challenge of the Defense Advanced Research Projects Agency (DARPA). ADS were competing first under tough environmental [51] and later in urban conditions.

There are several examples of implementing supervised driving. Paid safety drivers are supervising ADS performance on selected roads [24]. Customers use ADS in their daily trips and supervise its performance [25, 46]. ADS are driving by themselves in selected areas, but can be turned off if needed [30].

This work presented ideas for evaluating ADS with black box components. Understanding the mechanisms inside the black box could provide insight and build trust in the general abilities of ADS instead of showing that individual use cases can be handled. If we understand the abilities of ADS to correctly perceive road users, relevant infrastructure and objects then this would increase confidence in ADS to deal with possible traffic challenges. This knowledge is needed to avoid testing ADS in any situation, but fewer selected traffic situations. The challenge is to identify the “core-capabilities” necessary to handle all kinds of situations for ADS and to develop new metrics to test them.

4.4 Applying behaviour based methods

Analogous to the prerequisites of granting a driving license to human drivers, ADS can be evaluated with the presented methods. The method of experimental psychology reveals ADS’ abilities and can be seen as indicator of its level of performance. This is comparable to a human learning how to participate in traffic as a road user before even taking any driving lessons. Only when it reaches a certain level it is allowed to be observed in real traffic. Penetration testing can identify vulnerabilities. ADS need to be properly secured to reduce the risk of crashes due to misperception or malicious attacks. This method has no direct equivalent in the human driving test. It seems that it is mostly assumed that drivers are

able to assess situations accurately and deescalate if they are finding themselves in a challenging situation. The shadowing method can be seen as theoretical driving exam. It shows if ADS are capable of driving a vehicle. Supervised driving could be applied like a practical exam. Only if ADS pass actual driving on their own under natural traffic conditions are they allowed to do so without supervision.

The suggested approaches are behaviour based to mitigate the fact that the source code does not indicate the abilities of ADS. The end goal would be regulatory testing to validate ADS and issue approval for road use or not. It is out of the scope of this paper to describe this process in detail since behavioural based ADS testing is in a very early stage. Experimental studies are needed to test if these methods are actually useful in evaluating ADS. A first draft of integration into existing processes could be as follows. The method of experimental psychology can help in building a knowledge base of the capabilities of ADS. The actual studies can be performed by public or private research agencies. As long as the findings are widely accessible and the possibility of replication is ensured regulatory agencies can consider the knowledge in assessing ADS. This would be best achieved by following the guidelines of Open Science [38] and Open Access [47] to maximise public review and quality of the research.

The same procedure could be applied for penetration testing. In this case the requirements should include public and open contests to identify vulnerabilities of ADS. More research is needed to be able to say exactly when the knowledge base is good enough to proceed with regulatory approval. While the first two methods can be applied independently of ADS manufacturer, the shadowing and supervised driving method need to be performed by the OEM since they control the vehicles. The regulatory agency could require a certain performance in shadowing before allowing supervised driving. The same could be required from supervised driving. It is still unclear which performance indicator is well suited for this task, but it should allow the appraisal of ADS behaviour in natural driving conditions. In this regard these methods differ from existing ones like the Federal Motor Vehicle Safety Standards (FMVSS) or the European New Car Assessment Programme (Euro NCAP). These are investigating performance in specific areas only. With ADS emerging that regulate the complete driving task there is a need for a tool framework that evaluates the whole driving behaviour. The challenge in assessing whole behaviour is to show that task like “recognising a stop sign” work reliably under a wide range of conditions. The presented methods can contribute to such a framework.

There are exceptions in which the suggested methods are not directly applicable. For example, the shadowing and supervised driving method both require human driver controls for their application. They may be suitable for classic passenger vehicles, but not for new concepts of autonomous shuttles that do not feature any human driving controls. A workaround might be to have early prototypes with driving controls and remove them later when ADS have demonstrated its abilities sufficiently.

Seeing a machine primarily from its behavioural perspective is a concept that was recently taken up by Rahwan et al. [42]. Due to the inability to predict the exact behaviour of DNNs from the source code, they suggest a new branch of science, machine behaviour, dedicated to study the behavioural impact of machines. The idea was received with enthusiasm [22, 40]. At the moment DNNs are mostly developed by computer scientists, engineers and roboticists who are experts in building machines but not necessarily in evaluating behaviour. To understand and predict the impact of self-learning algorithms experts in experimental methodology, collective behaviour, social theory and neuroscience are needed [42]. Psychologists typically are experts in these areas and could support evaluating machine behaviour. A behavioural evaluation of the driving skills of autonomous cars could be a first step towards a science of machine behaviour in the automotive area dedicated to achieve safer and more convenient travels.

5 Conclusions

At the moment humans are still the best drivers we know. Even though crashes are usually caused by humans [48], they also handle a lot of situations without causing crashes. ADS have to achieve at least human level driving performance before they can provide a benefit. This will take time due to the cost and the challenge to train a machine to manage all those situations.

Four ideas that could be used to evaluate ADS were presented. They are suitable for testing any kind of black box system including DNNs. The validation of ADS is a necessary step to allow ADS unsupervised operation of a vehicle in everyday traffic. The suggested steps can help appraising the safety of the systems and ease the way towards integration of ADS into everyday traffic. However, experimental research is needed to evaluate if the methods work. It is necessary creating a testing environment and evaluating the approaches and to work closely with responsible offices to investigate if these methods can be implemented in the regulatory approval or not.

Even if achieved one day, safe autonomous driving is only the first step to participate in traffic. There are more challenges that need to be mastered before ADS are able to fully integrate themselves into everyday

traffic. A main focus of current ADS seems to be to drive on streets without bumping into others. However this is only a part of driving. It does not cover proactively supporting other road users in their goals. Not always doing what is allowed but refraining from rights to accommodate the needs of other road users. Often road users find themselves in unfavourable situations. Refraining from rights to support others helps to avoid risky manoeuvres and keeping the traffic safe. Neither do ADS communicate with other non-machine road users. Reactions are slower than mutual understanding. If a dangerous situation can be avoided by understanding and negotiating each other's goals, no risk mitigation is necessary. There are ideas how to communicate with other autonomous vehicles [23, 26], but successful 'autonomous vehicle to walking/riding/driving human communication' is very rare. At the moment ADS development doesn't seem to include anticipating the interpretation of ones actions by other road users and avoiding of unusual behaviour to not confuse them (e.g. stopping too long even when having right of way, so other road users assume they were granted way of right and then starting to drive). There are examples to predict other road user's behaviour [53]. However the information basis for it seems to be based on distances and speed, which allows only very limited predictions compared to what sources of information humans utilize. Finally, current ADS seem to lack the ability to drive with foresight: Understanding motivations and goals of other road users enable humans to anticipate actions but also reactions to the own behaviour. There are promising steps in this direction [8]. However, at the moment this ability to understand the world by taking another perspective is still in its infancy.

Humans avoid many crashes by not only thinking for themselves, but constantly developing hypothesis about the intentions of other road users. This allows them to predict their behaviour and adapt their own actions accordingly. It requires a theory of mind. This is the ability to see the world from the other's perspective. Being able to predict actions of other road users plays a huge part in traffic safety. It will be interesting to see how ADS will master this challenge or if human drivers have to adapt to the shortcomings of the machines in the end.

Acknowledgements
The authors would like to thank Matthias Powelleit (Ingenieur- und Verkehrspsychologie, TU Braunschweig) for his insight about neural networks and machine learning and Daniel Heß (Institut für Verkehrssystemtechnik, DLR Braunschweig) for his insight about autonomous driving. Also we would like to thank Bastian Farkas (Advanced Driver Information Technology GmbH) for his input on penetration testing.

Journal policies
The authors are not aware of any issues relating to journal policies.

Content publication
The Content of this manuscript has not been published, or submitted for publication elsewhere.

Authors' contributions
Fabian Utesch and Alexander Brandies conceived the presented idea and developed the theory. Paulin Pekezou Foupi contributed the theoretical framework from DNN and artificial intelligence research. Caroline Schießl supervised the project. Fabian Utesch wrote the manuscript in consultation from all authors. All authors provided critical feedback and helped shape the research and manuscript. All authors have approved the manuscript submission.

Funding
Open access funding provided by Projekt DEAL.

Availability of data and materials
Not applicable.

Competing interests
The authors are not aware of any potential competing interests.

Received: 18 November 2019 Accepted: 17 July 2020
Published online: 29 July 2020

References
1. Alvarez, S. (2020). *Tesla returns to Pwn2Own hacking competition with Model 3 as target and prize*. Retrieved from https://www.teslarati.com/tesla-model-3-returns-to-pwn2own-hacking-competition/
2. Assael, Y. M., Shillingford, B., Whiteson, S., & de Freitas, N. (2016). *LipNet: End-to-end sentence-level Lipreading. CoRR, abs/1611.01599*.
3. Atkinson, R. C., & Shiffring, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence (Ed.), *The psychology of learning and motivation*, (2nd ed., pp. 89–195). New York: Academic Press.
4. Baddeley, A., & Hitch, G. (1974). Working memory. In G. A. Bower (Ed.), *The psychology of learning and motivation*, (pp. 47–89). New York: Academic Press.
5. Barnard, M. (2016). *Tesla & Google Disagree about LIDAR — Which is right?* Retrieved from Clean Technica website: https://cleantechnica.com/2016/07/29/tesla-google-disagree-lidar-right/
6. Barnard, Y., Utesch, F., van Nes, N., Eenink, R., & Baumann, M. (2016). The study design of UDRIVE: The naturalistic driving study across Europe for cars, trucks and scooters. *European Transport Research Review*, *8*(2), 1–10. https://doi.org/10.1007/s12544-016-0202-z.
7. Bau, D., Zhu, J.-Y., Strobelt, H., Zhou, B., Tenenbaum, J. B., Freeman, W. T., & Torralba, A. (2018). *GAN Dissection: Visualizing and Understanding Generative Adversarial Networks*. Retrieved from https://arxiv.org/abs/1811.10597
8. Bhuiyan, J. (2017). *Alphabet is training law enforcement on how to handle self-driving car crashes*. Retrieved from recode website: https://www.recode.net/2017/10/15/16472896/alphabet-waymo-self-driving-law-enforcement-crashes-accidents
9. Bojarski, M., Yeres, P., Choromanska, A., Choromanski, K., Firner, B., Jackel, L., & Muller, U. (2017). Explaining How a Deep Neural Network Trained with End-to-End Learning Steers a Car (CoRR abs/1704.07911).
10. Bower, G. H., & Winzenz, D. (1969). Group structure, coding, and memory for digit series. *Journal of Experimental Psychology*, *80*(2, Pt.2), 1–17. https://doi.org/10.1037/h0027249.
11. Byford, S. (2016). *AlphaGo beats lee se-dol again to take Google DeepMind challenge series*. Retrieved from The Verge website: https://www.theverge.com/2016/3/12/11210650/alphago-deepmind-go-match-3-result
12. CGTN America. (2017). GO champ Ke Jie talks about his match with Google's AI. Retrieved from https://america.cgtn.com/2017/05/28/go-champ-ke-jie-talks-about-his-match-with-googles-ai
13. Czarnecki, K. (2019). Software engineering for automated vehicles: Addressing the needs of cars that run on software and data. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Companion Proceedings (ICSE-Companion)*, (pp. 6–8).
14. DLR. (2019). *Forschungsprojekt PEGASUS*. Retrieved from https://www.pegasusprojekt.de/de/about-PEGASUS

15. Donath, A. (2018). *Tesla will Sicherheitssoftware freigeben*. Retrieved from https://www.golem.de/news/open-source-tesla-will-sicherheitssoftware-freigeben-1808-135968.html
16. Ebbinghaus, H. (1971). *Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie (Unveränd. u. ungekürzte Aufl.)*. Darmstadt: Wiss. Buchges.
17. Eenink, R., Barnard, Y., Baumann, M., Augros, X., & Utesch, F. (2014). *UDRIVE: the European naturalistic driving study (transport research arena)*. Paris.
18. Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., … Song, D. (2018). Robust physical-world attacks on deep learning visual classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* Retrieved from https://arxiv.org/abs/1707.08945.
19. Gabbatt, A. (2011). *IBM computer Watson wins jeopardy clash*. Retrieved from The Guardian website: https://www.theguardian.com/technology/2011/feb/17/ibm-computer-watson-wins-jeopardy
20. Galleguillos, C., & Belongie, S. (2010). Context based object categorization: A critical survey. *Computer Vision and Image Understanding*, *114*(6), 712–722. https://doi.org/10.1016/j.cviu.2010.02.004.
21. Golson, J. (2016). *Tesla's new autopilot will run in 'shadow mode' to prove that it's safer than human driving*. Retrieved from The Verge website: https://www.theverge.com/2016/10/19/13341194/tesla-autopilot-shadow-mode-autonomous-regulations
22. Hao, K. (2019). *AI researchers want to study AI the same way social scientists study humans*. Retrieved from https://www.technologyreview.com/s/613440/ai-researchers-want-to-study-ai-the-same-way-social-scientists-study-humans/
23. Heß, D., Löper, C., & Hesse, T. (2017). *Safe Cooperation of Automated Vehicles (AAET 2017)*.
24. Johnson, L. & Fitzsimmons, M. (2018). *Uber self-driving cars: everything you need to know*. Retrieved from https://www.techradar.com/news/uber-self-driving-cars
25. Karpathy, A. (2020). *AI for Full-Self Driving (Scaled Machine Learning Conference)*. Retrieved from https://youtu.be/hx7BXih7zx8
26. Kaup, M., Willrodt, J.-H., Schieben, A., & Wilbrink, M. (2019). *interACT D4.3 – Final design and HMI solutions for the interaction of AVs with user on-board and other traffic participants ready for final implementation. Retrieved from interACT project website*: https://www.interact-roadautomation.eu/projects-deliverables/
27. Kline, M. (2017). *Elon Musk's 'Dota 2' experiment is disrupting Esports in a big way*. Retrieved from Mashable website: https://mashable.com/2017/08/31/elon-musk-dota2-ai-experiment/#7Mu4MnCBxsOY
28. Krauthammer, C. (1997). *Be Afraid*. Retrieved from https://www.weeklystandard.com/be-afraid/article/9802
29. Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324. https://doi.org/10.1109/5.726791.
30. Locklear, M. (2018). *Waymo reportedly returns safety drivers to its autonomous cars*. Retrieved from https://www.engadget.com/2018/11/27/waymo-returns-safety-drivers-to-autonomous-cars/
31. Lubben, A. (2018). *Self-driving Uber killed a pedestrian as human safety driver watched*. Retrieved from https://www.vice.com/en_us/article/kzxq3y/self-driving-uber-killed-a-pedestrian-as-human-safety-driver-watched
32. Madrigal, A. C. (2017). *Inside Waymo's Secret World for Training Self-Driving Cars*. Retrieved from https://www.theatlantic.com/technology/archive/2017/08/inside-waymos-secret-testing-and-simulation-facilities/537648/
33. Maurer, M., Gerdes, J. C., Lenz, B., & Winner, H. (2015). *Autonomes Fahren*. Berlin: Springer Berlin Heidelberg.
34. Millard-Ball, A. (2019). The autonomous vehicle parking problem. *Transport Policy*, *75*, 99–108. https://doi.org/10.1016/j.tranpol.2019.01.003.
35. Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97. https://doi.org/10.1037/h0043158.
36. Munroe, R. (2011). *XKCD 936: Password strength*. Retrieved from XKCD website: https://xkcd.com/936/
37. Murre, J. M. J., & Dros, J. (2015). Replication and analysis of Ebbinghaus' forgetting curve. *PloS one*, *10*(7), e0120644. https://doi.org/10.1371/journal.pone.0120644.
38. Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., … Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374.
39. O'Kane, S. (2018). How Tesla and Waymo are tackling a major problem for self-driving cars: Data. Retrieved from The Verge website: https://www.theverge.com/transportation/2018/4/19/17204044/tesla-waymo-self-driving-car-data-simulation
40. Pavlus, J. (2019). *The Anthropologist of Artificial Intelligence*. Retrieved from https://www.quantamagazine.org/iyad-rahwan-is-the-anthropologist-of-artificial-intelligence-20190826/
41. Pei, K., Cao, Y., Yang, J., & Jana, S. (2017). *DeepXplore: Automated Whitebox Testing of Deep Learning Systems (SOSP '17). Shanghai, China*.
42. Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., … Wellman, M. (2019). Machine behaviour. *Nature*, *568*(7753), 477–486. https://doi.org/10.1038/s41586-019-1138-y.
43. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *Why should I trust you?: Explaining the predictions of any classifier. CoRR, abs/1602.04938*.
44. Ross, Philip, E. (2016). *Tesla reveals its Crowdsourced autopilot data*. Retrieved from IEEE Spectrum website: https://spectrum.ieee.org/cars-that-think/transportation/self-driving/tesla-reveals-its-crowdsourced-autopilot-data
45. SAE. (2014). *Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems*. Retrieved from https://www.sae.org/standards/content/j3016_201401/
46. Schmidt, B. (2019). *Over half a million Tesla electric cars have Autopilot 2 – what's the big deal?* Retrieved from https://thedriven.io/2019/07/08/over-half-a-million-tesla-electric-cars-have-autopilot-2-whats-the-deal/
47. SPARC and PLOS. (2012). *HowOpenIsIt? Open Access Spectrum*. Retrieved from https://sparcopen.org/our-work/howopenisit/
48. Statistisches Bundesamt. (2019). *Verkehrsunfälle 2018 (Fachserie 8, Reihe 7)*.
49. Steiniger, U., Schöner, H.-P., & Schiementz, M. (2015). *Anforderungen an Werkzeuge zur Bewertung und Absicherung von Systemen zum assistierten und automatisierten Fahren (7. Tagung Fahrerassistenz: Der Weg zum automatischen Fahren). München*.
50. Tesla. (2018). *Tesla product security*. Retrieved from tesla.com/about/security
51. Thrun, S., Montemerlo, M., Dahlkamp, H., Stavens, D., Aron, A., Diebel, J., … Mahoney, P. (2006). Stanley: The robot that won the DARPA grand challenge. *Journal of Field Robotics*, *23*(9), 661–692. https://doi.org/10.1002/rob.20147.
52. Tian, Y., Pei, K., Jana, S., & Ray, B. (2018). DeepTest: Automated testing of deep-neural-network-driven autonomous cars. In *ICSE '18, proceedings of the 40th international conference on software engineering*, (pp. 303–314). New York: Association for Computing Machinery. Retrieved from. https://doi.org/10.1145/3180155.3180220.
53. Ulbrich, S., & Maurer, M. (2015). Towards tactical lane change behavior planning for automated vehicles. In *2015 IEEE 18th International Conference on Intelligent Transportation Systems - (ITSC 2015)*, (pp. 989–995).
54. Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., … Silver, D. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, *575*(7782), 350–354. https://doi.org/10.1038/s41586-019-1724-z.
55. Vlasic, B. (2016). *Self-driving Tesla was involved in fatal crash, U.S. Says*. Retrieved from https://www.nytimes.com/2016/07/01/business/self-driving-tesla-fatal-crash-investigation.html
56. Zhang, Q., Nian Wu, Y., & Zhu, S.-C. (2018). Interpretable convolutional neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Publisher's Note