More than a decade ago, when the idea of self-driving cars still seemed a long way from reality, an experiment revealed how odd machine-made decisions can seem to the casual observer.

In the Urban Challenge competition organised by US research agency, DARPA, two autonomous vehicles collided at low speed in a way that, at first, baffled the designers.

Cornell University’s Skynet had trouble with its software and slowed almost to a halt – but not completely. It was still lurching forward intermittently. Approaching from behind, MIT’s Talos considered everything moving below a certain speed to be effectively static and cut in front – just at the point Skynet lurched forward again, this time hitting Talos.

More than a decade on, a misclassification problem looks to have played a major role in the death of a cyclist that crossed a dark road in Arizona in front of one of Uber’s experimental self-driving cars. The information reported anonymous sources who claimed the car’s software registered an object in front of the vehicle, but treated it in the same way it would a plastic bag or tumbleweed carried on the wind.

Autonomous vehicles have to make split-second decisions based on how they classify the objects in the scene in front of them. Developers currently favour using AI techniques because purely algorithmic methods are likely to be too complex to develop. But when AI makes mistakes, working out why becomes the problem. It is an issue that extends way beyond self-driving vehicles.

Pinning down the misclassification to a flaw in the training data or the structure of the learning algorithm itself is much harder than understanding that something has gone wrong. Despite the huge amount of R&D money now being poured into deep learning in particular, the ability to validate and verify the systems remains poor.

Researchers regard deep neural networks as being, for the most part, black boxes.

Professor Philipp Slusallek, scientific director at the German Research Centre for AI (DFKI), said in his keynote at CDNLive EMEA 2018: “At the moment, we don’t have any verification technology for deep-learning networks. We don’t yet understand the limits of this technology.”

Rich Caruana, senior researcher at Microsoft Research, says AI today has a core problem: “It’s an idiot savant. It has an incredibly narrow understanding of the world.”

Deep neural networks demonstrate this in the way they sometimes home in on features in an image that are completely different to those used by humans. Often, unconscious choices in training data will give too much significance to elements that, to the machine will seem important, but which are really just the result of coincidences.

A system to classify images as pets might well use the presence of a cardboard box to push the conclusion in the direction of “cat” simply because many are photographed
at the Data, Learning and Inference Co-Founder and CTO, Amnon Shashua, powered car unacceptable. Mobileye in the real world and make the AI - no human would."

As they don’t fall into the same traps different. It gives them some strength we think. So the traps they fall into are think is still very different from the way showed several years ago.

The conditions that trip up a deep neural network are, frequently, far less obvious. Researchers working at Kyushu University in Japan showed last autumn how changing just one pixel within an image could cause a network to misclassify a picture of an animal. Neural networks can often find hidden patterns in noise, as a team from the University of Wisconsin showed several years ago.

Caruana argues: “The way they think is still very different from the way we think. So the traps they fall into are different. It gives them some strength as they don’t fall into the same traps we do, but they also fall into traps that no human would.”

Such traps could severely damage the user acceptance of systems in the real world and make the AI-powered car unacceptable. Mobileye Co-Founder and CTO, Amnon Shashua, at the Data, Learning and Inference (DALI) conference in early April, warned of a possible winter for advanced driver assistance systems that mirrors the many times AI has been frozen out of the technological mainstream because it failed to deliver on its promises.

“Imagine you are driving a car and the car suddenly brakes for no apparent reason. You will take the car back to the dealership. We are talking about designing things that have to show zero false positives. If you go to fully autonomous driving it gets even worse.”

There are other problems besides braking or not braking at the wrong times. Shashua points out autonomous cars will need to work out how to handle tasks such as merging onto busy motorways – and understand when other drivers will make space for them – or be left stranded at the side of the road. Making the robot understand the subtle behavioural and environmental clues that affect real-world driving will take much more work, due to the way neural networks currently operate.

“Deep learning needs lots and lots of data,” says Slusallek. It could take millions of miles of real-world testing to reveal enough combinations of conditions to both train and validate the AI’s behaviour. “The only way we can make sure our systems do the right thing is to generate the data synthetically,” Slusallek adds, so that the systems can be tested much more extensively in a virtual space.

XAI programme
Extensive behavioural testing can only go so far. Engineers need to be sure that the system is not interpreting incoming data incorrectly and has simply been lucky in not triggering unwanted reactions. Caruana argues for AI to be explainable, and has pursued research on alternatives to deep learning that are easier to interpret. DARPA is equally keen to find AI systems that are explainable. The XAI programme got underway two years ago and is recruiting researchers from the US and beyond.

Subramanian Ramamoorthy and colleagues from the University of Edinburgh, for example, are working with the Palo Alto Research Center (PARC) on one project under the XAI banner to infer models from analysis of the behaviour of AI systems.

Researchers working outside the DARPA programme are pursuing a number of ways to address AI verification. Xiaowei Huang and colleagues from the University of Oxford have worked on formal verification techniques that perturb the network and its input data to determine the stability of classifications made by the network. Runtime performance is a major challenge, but work along these lines should make neural networks more resistant to noise-injection attacks as well as poorly balanced training sets.

Other teams are focusing on supplementing or replacing black-box AI with models that lend themselves more readily to analysis. University of Toronto researcher, Professor Angela Schoellig, is working on coupling programmatic models with machine learning. This should make it possible to start with a system that behaves safely, because it is driven by the physics models, but which gradually incorporates more behaviour that is learned when it performs.

A team from the University of California at Berkeley aims to use a second AI system based on Bayesian learning as a back-up for the more sophisticated deep-learning core. In principle, this would force the robot to make safer choices and not be misled by anomalies in the training.

At such an early stage in the development of dependable AI systems, the way forward remains unclear. But the need for machine learning in systems that are expected to be safe under all conditions is forcing much deeper investigation of how these technologies really work and how they make mistakes.