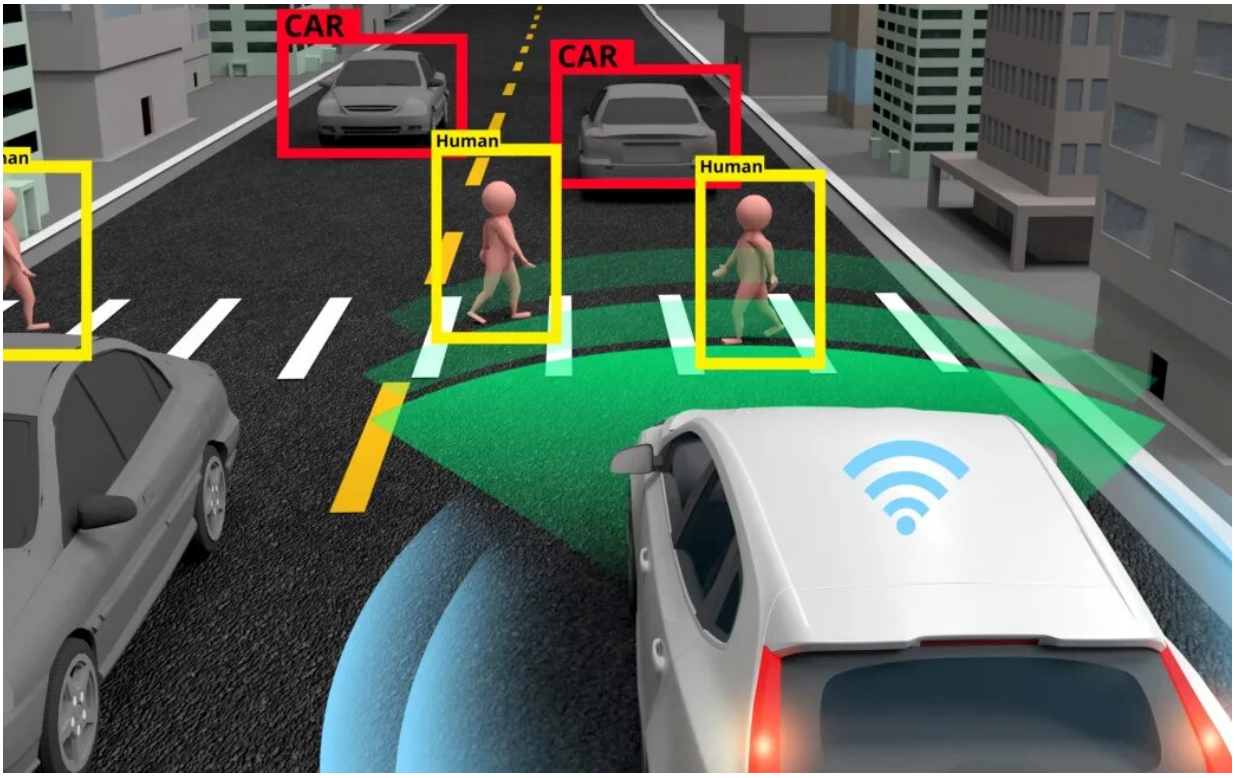
Study: Machine learning models cannot be trusted with absolute certainty

April 18 2023, by Paavo Ihalainen



Self-driving cars are expected to have better safety measures than human drivers. The more critical the system is, the more relevant is the capacity for minimizing the associated risks. Credit: Akarat Phasura

An article titled "On misbehaviour and fault tolerance in machine learning systems," by doctoral researcher Lalli Myllyaho was named one

of the best papers in 2022 by the *Journal of Systems and Software*.

"The fundamental idea of the study is that if you put critical systems in the hands of artificial intelligence and algorithms, you should also learn to prepare for their failure," Myllyaho says.

It may not necessarily be dangerous if a streaming service suggests uninteresting options to users, but such behavior undermines trust in the functionality of the system. However, faults in more critical systems that rely on machine learning can be much more harmful.

"I wanted to investigate how to prepare for, for example, computer vision misidentifying things. For instance, in computed tomography artificial intelligence can identify objects in sections. If errors occur, it raises questions about to what extent computers should be trusted in such matters, and when to ask a human to take a look," says Myllyaho.

The more critical the system is, the more relevant is the capacity for minimizing the associated risks.

More complex systems generate increasingly complex errors

In addition to Myllyaho, the study was carried out by Mikko Raatikainen, Tomi Männistö, Jukka K. Nurminen and Tommi Mikkonen. The publication is structured around expert interviews.

"Software architects were interviewed about the defects and inaccuracies in and around machine learning models. And we also wanted to find out which design choices could be made to prevent faults," Myllyaho says.

Should machine learning models contain broken data, the problem can

extend to systems in whose implementation the models have been used. It is also necessary to determine which mechanisms are suited to correcting errors.

"The structures must be designed to prevent radical errors from escalating. Ultimately, the severity to which the problem can progress depends on the system."

For example, it is easy for people to understand that with autonomous vehicles, the system requires various safety and security mechanisms. This also applies to other AI solutions that need appropriately functioning safe modes.

"We have to investigate how to ensure that, in a range of circumstances, artificial intelligence functions as it should, that is with human rationality. The most appropriate solution is not always self-evident, and developers must make choices on what to do when you cannot be certain about it."

Myllyaho has expanded on the study by developing a related mechanism for identifying faults, although it has not yet advanced to an actual algorithm.

"It's just an idea of neural networks. A functional machine learning model would be able to switch working models on the fly if the current one does not work. In other words, it should also be able to predict errors, or to recognize indications of errors."

Recently, Myllyaho has concentrated on finalizing his doctoral thesis, which is why he is unable to say anything about his future in the project. The IVVES project headed by Jukka K. Nurminen will continue to carry out its work in testing the safety of machine learning systems.

More information: Lalli Myllyaho et al, On misbehaviour and fault tolerance in machine learning systems, *Journal of Systems and Software* (2022). DOI: 10.1016/j.jss.2021.111096