ABSTRACT

Autonomous driving needs machine learning, because it relies so heavily on perception. But machine learning is notoriously unpredictable and unverifiable. How then can an autonomous car ever be convincingly safe? Dr. Jackson and his research team have been exploring the classic idea of a runtime monitor: a small trusted base that executes in parallel and intervenes to prevent violations of safety.

Unfortunately, in this context, the traditional runtime monitor is not very plausible. If it processes sensor data itself, it is likely either to be no less complex than the main system, or to be too crude to allow early intervention. And if it does not process sensor data, and instead relies on the main system for that, the key benefit of a small trusted base is lost.

The research team has been pursuing a new approach in which the main controller constructs a “certificate” that embodies a run-time safety case. The monitor is only responsible for checking the certificate, which gives the desired reduction in complexity, exploiting the typical gap between the cost of finding solutions to computational problems and the cost of checking them.

Dr. Jackson will illustrate this idea with some examples his team has implemented in simulation, with the disclaimer that this research is in the early stages. His hope is to provoke an interesting discussion.

BIO

Daniel Jackson is a Professor of Computer Science at MIT, a MacVicar teaching fellow, and an Associate Director of the Computer Science and Artificial Intelligence Laboratory. His research has focused primarily on software modeling and design. Jackson is also a photographer; his most recent projects are Portraits of Resilience (http://portraitsofresilience.com), and At a Distance (https://dnj.photo/projects/distance). His book about software design, The Essence of Software: Why Concepts Matter for Great Design, will be published this fall by Princeton University Press.