2.1 System-level verification

Practical guidance – cross-domain and automotive

Author: ATM project at Australia National University

Overview of approach

As robotics and autonomous systems (RAS) start to become consumer products, questions arise on how to help consumers compare the safety of different RAS. This is the question that is addressed in this guidance.

Most work on safety of RAS focuses on the developers' point of view. Here we discuss safety indicators and assessment mechanisms for consumers. This aim implies that the safety indicator must be simple yet informative, while the assessment mechanism must be sufficiently fast. Specifically, we aim to develop a mechanism that can automatically assess the safety of a RAS as a holistic system, with minimal reliance on information about the inner working of the RAS, and present this assessment in a user-friendly manner. In this guidance we will focus on assessing catastrophic accidents.

The proposed safety indicator and assessment are akin to the well-known NCAP safety rating for cars. However, the mechanism that generates the NCAP rating requires substantial modifications for RAS. NCAP rating has been designed to test the safety of the car's mechanics and hardware, whose behaviour generally changes gradually and very slowly – in the order of multiple years. In contrast, most of today's RAS are run by software that automatically adapts their behaviour – often in the order of days – as more data are gathered and regular patches/updates are applied. Although these adaptations are supposed to improve performance, it is often unclear as to their effects on the fringe rare cases where catastrophic accidents often happen.

Euro-NCAP has started to assess the autonomy augmentation of new car models. However, these assessments still follow a rather ad-hoc mechanism (e.g. a mannequin is being "dragged around" following a certain fixed trajectory to test autonomous collision avoidance capabilities). The problem is that such an ad-hoc mechanism is easy to fool by a RAS once the sets of test scenarios are known.

To better test the fast-adaptable nature of RAS, we propose an automatic testing mechanism that deliberately finds human behaviours that cause catastrophic accidents. The similarity between behaviours that cause catastrophic accidents and those of common human behaviour when interacting with the RAS indicates how likely catastrophic accidents happen, and can be used to compute the safety rating of the RAS.

The overall assessment mechanism itself is illustrated in Figure 1. Given a clone of the RAS software, or the RAS itself, our proposed assessment mechanism will compute a set of safe and kamikaze trajectories closest to those safe trajectories. The safety indicator, SKD, is then the average distance between samples of pairs of safe and its closest kamikaze trajectories. This SKD provides an upper bound on the probability that a small deformation changes a collision-free trajectory of the adversary into a colliding one.

Full technical details of the approach are available through the project's website. The software relating to this approach is available to download.

Note that the guidelines are based on a proposed testing mechanism that is still at its infancy. Therefore, these guidelines will need to undergo an extensive review and adjustment prior to deployment, as the testing mechanism matures.

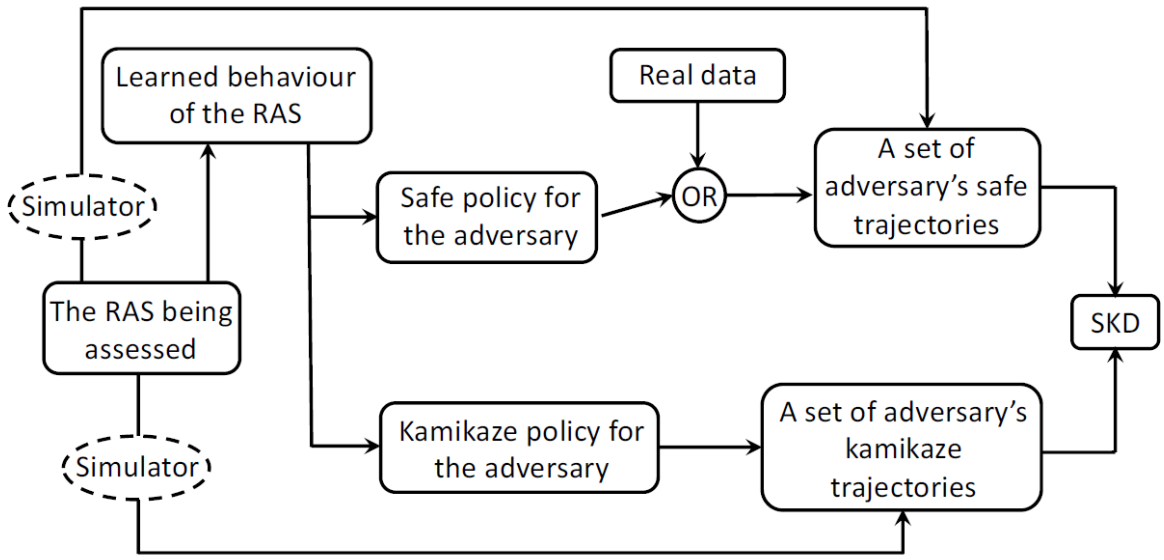

Figure 1 - Proposed mechanism to assess the safety of a RAS. Dashed ellipse means it may or may not be used.

Different stakeholders require different levels of knowledge and can consider parts of the mechanism as a black box. Figure 2 illustrates an overview of the level of knowledge requires for the different stakeholders with respect to the Bloom's Taxonomy.

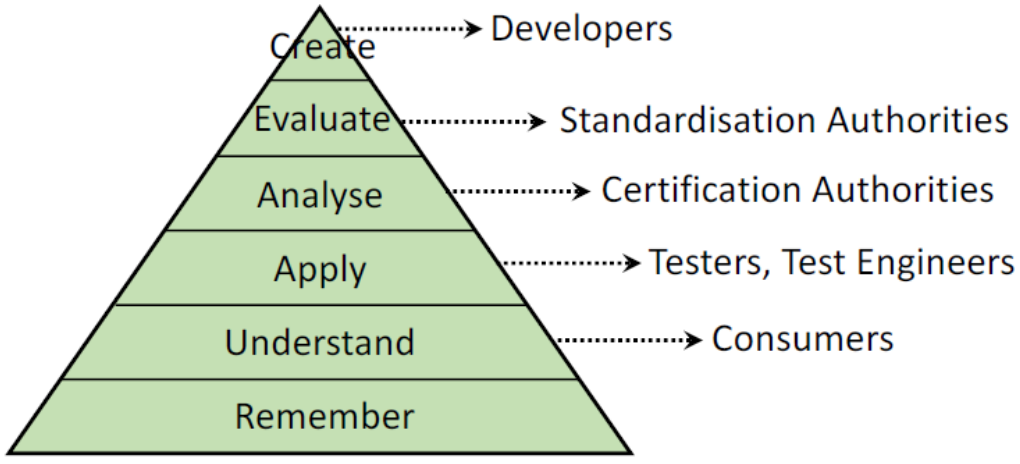

Figure 2 - Required knowledge levels for the different stakeholders, with respect to the Bloom's Taxonomy

Details of the necessary knowledge required by the different stakeholders are described in the subsequent subsections:

- Consumers of RAS
- Testers
- Testing engineers

- Certification authorities
- Standardisation authorities
- Developers of RAS

Consumers of RAS

Given the frequent updates and patches that may change the behaviour of RAS, assessment of its safety must be performed relatively often, at the very least, after every major update is performed. To encourage consumers of RAS to maintain such a frequent safety assessment, it is important to make the assessment mechanism as convenient as possible to the consumers. Therefore, we propose a mechanism akin to today's car wash, where consumers can bring their RAS to a RAS safety assessment service, leave it for 15–30 minutes and come back with the result of the assessment. Alternatives, such as uploading a clone of the RAS' software, rather than bringing the physical RAS to an assessment centre, may be possible too. Regardless of the exact assessment mechanism, we believe minimising the RAS down time due to safety assessment would be important for the adoption of frequent safety assessment of RAS.

We believe it is sufficient for consumers to know what the safety indicator SKD means, to enable them to compare the safety of one model of RAS and another. They do not need to understand how the SKD is computed. This is similar to today's car's consumers: they can differentiate the safety of different car models via the NCAP rating of the car, but do not need to understand the details on how a car receives its NCAP rating. This is illustrated in Figure 3.

We foresee that some consumers may want to perform the safety assessment themselves. Therefore, we believe it will be beneficial to develop a DIY-kit for such consumers. However, these consumers must have the proficiency at the level of testers (as described in subsequent subsection). Moreover, for DIY safety assessment results, we suggest a stringent certification process to be required for RAS that operates mostly in environments cohabited by humans.

Given the above requirements, consumers of RAS who use the safety assessment service only need to be informed about the safety indicators and what it means, no new training is required for them. However, for those intending to perform the assessment themselves, training and understanding of the safety indicator to the level required of a tester is necessary.

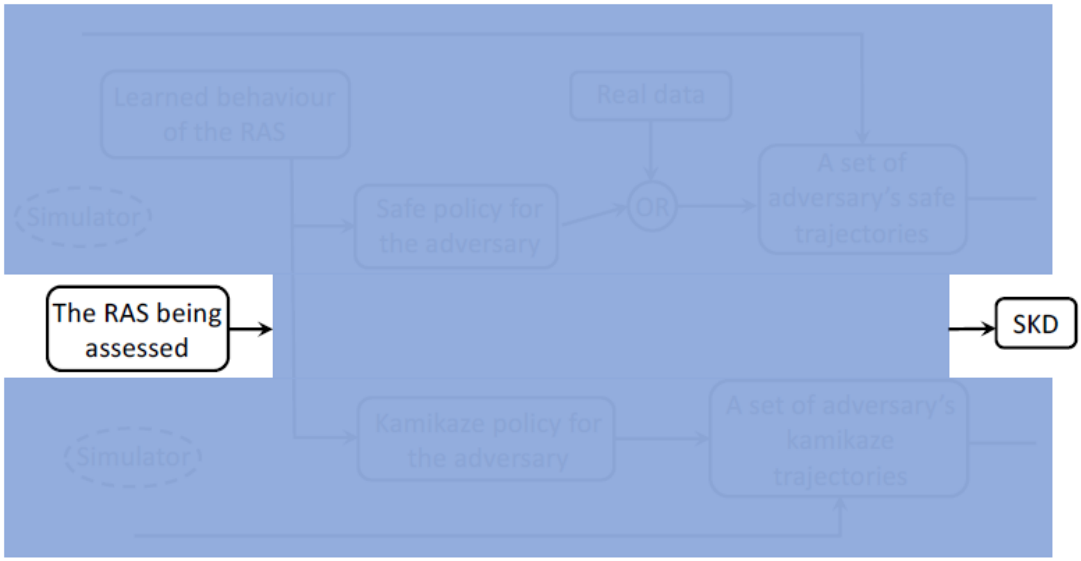

Figure 3 - Consumers can view the mechanism shown in Figure 1 as a black box. The greyed area does not need to be revealed to the consumer.

Testers

For providers of the safety assessment service, they need to be able to properly apply the software to compute SKD for the car being assessed. The software can be designed, such that the parameters are fixed, and therefore the tester does not need to know the inner working of the safety and kamikaze trajectories generation nor the SKD computation.

However, since SKD relies on samples of safe and kamikaze trajectories, the tester needs to ensure that the number of samples used are sufficient. Different systems may require different number of samples to achieve statistical significance. Based on the variance of the results when the default number of samples are used, one can compute the number of samples required to achieve statistically significant results. In addition to the number of samples, obviously, a tester must be able to clone the software of RAS, so that it can be provided as input to the assessment software, or to run tests on the physical RAS.

This is illustrated in Figure 4 where the greyed area does not need to be revealed to the testers, while the greyed boxes indicate that the testers need to know these components exists and some of the input/output to these components.

The above requirement implies that the tester needs to have a good understanding of statistics and probability, computer system, and high-level programming. Typical courses given to second year undergraduate in most natural sciences, computing, and engineering program would suffice.

A technical high school plus a one-year intensive training may be sufficient to become a tester.

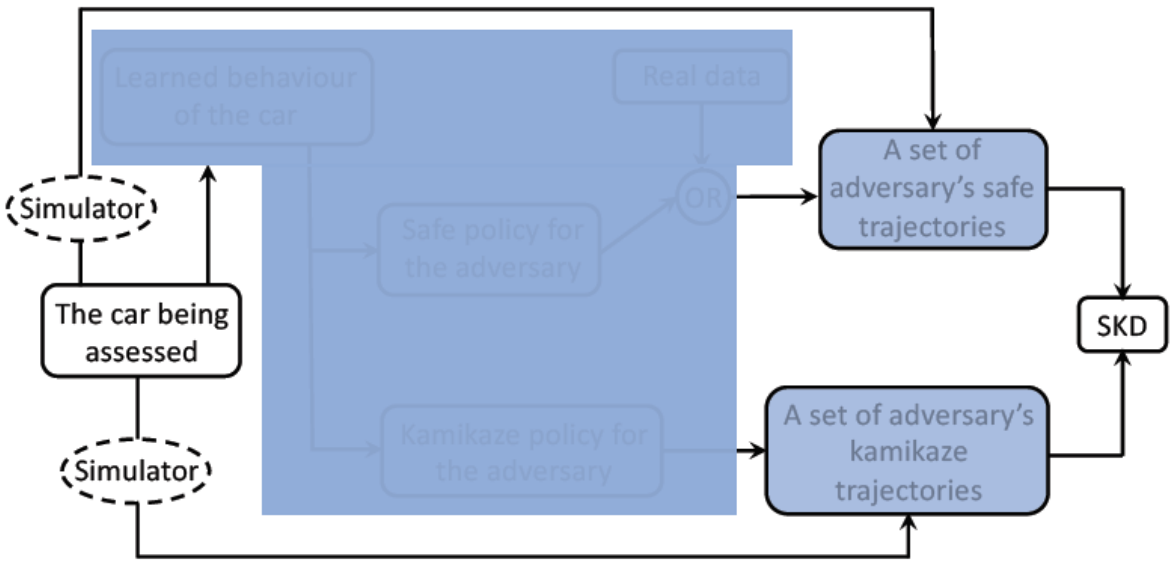

Figure 4 - Testers need to know more than the consumers, but not the entire mechanism.

The greyed area does not need to be revealed to the testers, while the greyed boxes indicate that the testers need to know these components exists and some of the input/output to these components.

Testing engineers

In addition to testers, a safety assessment service would require a testing engineer to maintain the software for testing. We foresee that parameters of the testing mechanism and its components may need to be adjusted, for instance, based on new regulations or different types of RAS being assessed. Let's call such parameters as user-modifiable parameters. The exact parameters included as user-modifiable parameters must be approved by the certification authorities, and such adjustments should be made via a configuration file, without changing the programs. Despite these measures, adjustments of the parameters require a testing engineer who understands the details of testing mechanism to the level that they understand the effects of adjusting a user-modifiable parameter to the resulting safety indicator. This is illustrated in Figure 5 where the greyed boxes indicate that the test engineers need to know these components exists and some of the input/output to these components.

The above requirement means the testing engineer needs to have the knowledge that a tester has, plus a good understanding of the technical methods being used by the testing mechanism. They will need to be well trained in artificial intelligence and machine learning techniques used by the testing mechanism, as well as in applied statistics and probability as necessary to understand the components of the testing mechanism.

A testing engineer would likely require training equivalent to a three-year diploma.

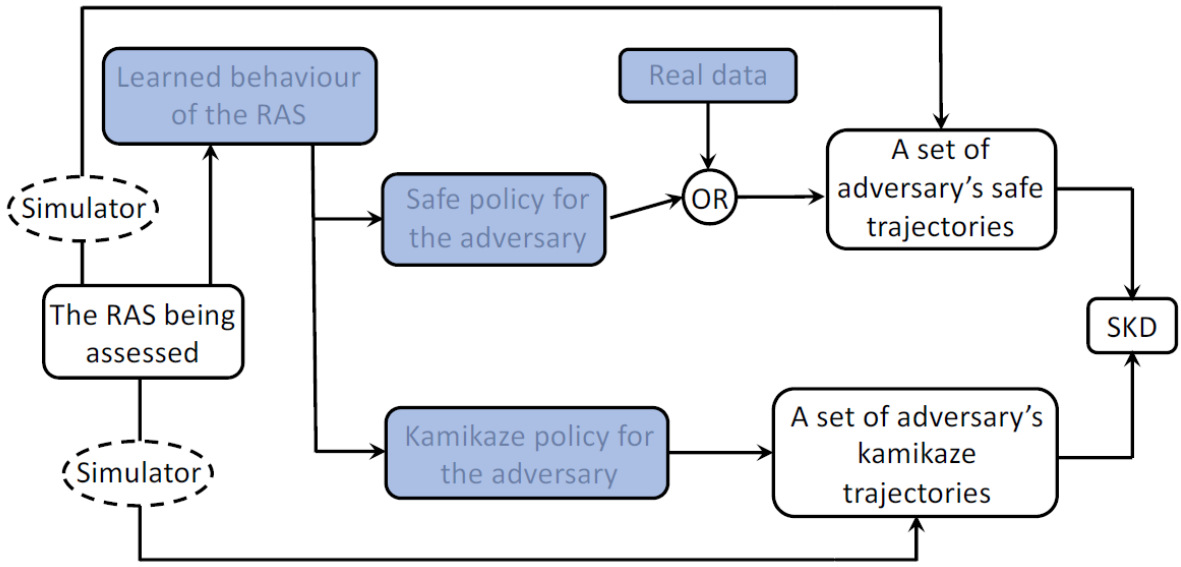

Figure 5 - Testing engineers need to know the components of the assessment mechanism, but not the details of each and every component. The greyed boxes indicate that the test engineers need to know these components exists and some of the input/output to these components.

Certification authorities

We suggest two types of certification:

1. Certification of the resulting safety indicator
2. Certification of the assessment software.

The first type of certification will ensure that the parameters are set appropriately. This certification can be performed automatically by augmenting the safety assessment software.

Certification of the assessment software and safety indicator certifier will ensure that the software has not been altered, intentionally or unintentionally. We require that an acceptable safety indicator must be generated and certified by a certified software. This certification can be performed on a regular but less frequent basis (e.g. once a year).

The above certification procedure will likely avoid delays in the usability of the safety indicator due to certification bottlenecks. Such a bottleneck will cause inconveniences to the RAS consumer, and result in less adoption of the safety assessment, and therefore should be avoided.

To perform the software certification, a certifier must understand the concepts, techniques, and software used for the assessment. They will need to have the knowledge of a Testing Engineer, plus a more in-depth understanding on how the concepts and methods used by the assessment mechanism are being implemented. Therefore, in addition to artificial intelligence, machine learning, and applied statistics and probability, a certifier must be well trained in software engineering too.

A certification authority would likely require training equivalent to an undergraduate degree.

Standardisation authorities

A safety indicator will eventually be used to compare different RAS models in the market. To ensure comparability, there needs to be standardisation on the acceptable safe trajectories and kamikaze trajectories generation, as well as the acceptable statistical confidence levels of the safety indicator. These standardisation needs to be based on extensive safety indicator results on multiple RAS.

Standardisation authorities need to possess the knowledge of a testing engineer plus an in-depth understanding of applied statistics and probability, to properly identify suitable parameters to be used.

Someone who sets standard for the safety assessment mechanism would likely require training equivalent to an undergraduate degree.

Developers

Both RAS developers and the safety assessment developers must understand the technical details of the safety assessment mechanism, so as to improve the safety of their RAS and the assessment mechanism. The developers would need to have in-depth conceptual and technical understanding on the safety assessment approach and methods, as well as the nuances of sufficient statistics and comparability of the safety indicators.

These developers could be separated into engineers, who implement the concepts and technical methods, and researchers, who design the next generation RAS and safety assessment mechanism. The level of education required for them would be similar to those require of RAS engineers and researchers today.