Summarising a Framework for the Certification of Reliable Autonomous Systems

Fisher, Michael; Schlingloff, Bernd-Holger; Mascardi, Viviana; Winikoff, M.D.; Rozier, Kristin Yvonne; Yorke-Smith, N.

Publication date
2021

Document Version
Final published version

Citation (APA)
Fisher, M., Schlingloff, B.-H., Mascardi, V., Winikoff, M. D., Rozier, K. Y., & Yorke-Smith, N. (2021). *Summarising a Framework for the Certification of Reliable Autonomous Systems*. 1733-1734. Abstract from 20th International Conference on Autonomous Agentsand Multiagent Systems. http://www.ifaamas.org/Proceedings/aamas2021/pdfs/p1733.pdf

Summarising a Framework for the Certification of Reliable Autonomous Systems

JAAMAS Track

Michael Fisher
University of Manchester, UK
michael.fisher@manchester.ac.uk

Viviana Mascardi
University of Genova, IT
viviana.mascardi@unige.it

Kristin Y. Rozier
Iowa State University, US
kyrozier@iastate.edu

Bernd-Holger Schlingloff
Humboldt University and FOKUS, DE
hs@informatik.hu-berlin.de

Michael Winikoff
Victoria University of Wellington, NZ
michael.winikoff@vuw.ac.nz

Neil Yorke-Smith
Delft University of Technology, NL
n.yorke-smith@tudelft.nl

ABSTRACT

This extended abstract summarises the contributions from the journal article Fisher et al. [2].

KEYWORDS

Autonomous Systems; Reliability; Verification

ACM Reference Format:
Michael Fisher, Viviana Mascardi, Kristin Y. Rozier, Bernd-Holger Schlingloff, Michael Winikoff, and Neil Yorke-Smith. 2021. Summarising a Framework for the Certification of Reliable Autonomous Systems: JAAMAS Track. In *Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), Online, May 3–7, 2021,* IFAAMAS, 2 pages.

1 PROBLEM

Increasingly we are delegating responsibilities to software, both in terms of digital interactions and practical situations. By delegating control over (even a few) safety critical functions we are relying on software for our safety, security or privacy. In parallel these autonomous systems – which make decisions and potentially take actions on their own – are also becoming sophisticated and complex. We have reached the stage that few users understand exactly how these systems work and so cannot assess whether they will be reliable and trustworthy. In particular, the ways in which autonomous decisions are made may be opaque: hence, not only do users not know what decisions will be taken, but they have little idea *why* those decisions are selected.

Although this delegation of responsibility appears problematic, it is the price we pay for autonomy. Further, as these systems become more sophisticated, we are left with no choice but to delegate further key responsibilities. For example, if we want a car to drive by itself, we can no longer just delegate the control of very basic operations (such as speed/lane control), but must also delegate to it the requirement to follow the road traffic rules. Let us follow this example further. Even if the vehicle has been designed (and verified) to follow all road traffic rules, what will it do in unexpected or unanticipated situations? We quickly end up with moral–ethical decisions that we would ideally want the human 'driver' to intervene on. But, again, fully autonomous vehicles may allow human drivers to become detached, lose concentration, and lose situational awareness (perhaps even fall asleep). What will the vehicle do then? All these aspects come in to play when designing, verifying and, ultimately, certifying autonomous systems.

In Fisher et al. [2] we tackle the analysis of true autonomy by:

(1) proposing a framework for viewing (and indeed building) autonomous systems in terms of three layers;
(2) showing that this framework is general, by illustrating its application to a range of systems, in a range of domains;
(3) discussing how certification/regulation might be achieved, breaking it down by the three layers; and
(4) articulating a range of challenges and future work, including challenges to regulators, to researchers, and to developers.

This type of work is necessary for several reasons. Most importantly, at the time of writing there are no standards/regulations (apart from BS8611 [1]) that consider truly autonomous systems. Those that mention such aspects appear to assume that either there will always be a human operator/driver/pilot who can quickly resolve any conflicts or that systems really have no ability to make their own decisions in unanticipated situations.

2 CONTEXT

Any autonomous systems to be deployed need to pass regulatory barriers, which in turn appeal to standards concerning system behaviour and functions. Since autonomy can be relevant across a range of different sectors there is a vast range of standards from all across the leading standardisation organisations, such as CENELEC, IEC, IEEE, and ISO. We discuss many of these in Fisher et al. [2], across sectors including air, rail, robotics, and software, but note that issues concerning truly autonomous systems have yet to make it into these standards. Current standards and regulations are not ready to cope with fully autonomous systems that may raise safety issues, thus motivating our use of stronger formal processes.

We also discuss the related problem of formalising specifications, whether from standards themselves or from user requirements, and explore the general issues around 'autonomy' and 'uncertainty'. For example, we will sometimes need to consider not just *what* an autonomous system does, but also *why* it chose to do it. For instance, there is a difference between a car breaking the speed limit because it has an incorrect belief about the speed limit, and a car going too fast because it believes that temporarily speeding is the best, or even the only, way to avoid an accident. Such assessment of

Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), U. Endriss, A. Nowé, F. Dignum, A. Lomuscio (eds.), May 3–7, 2021, Online.

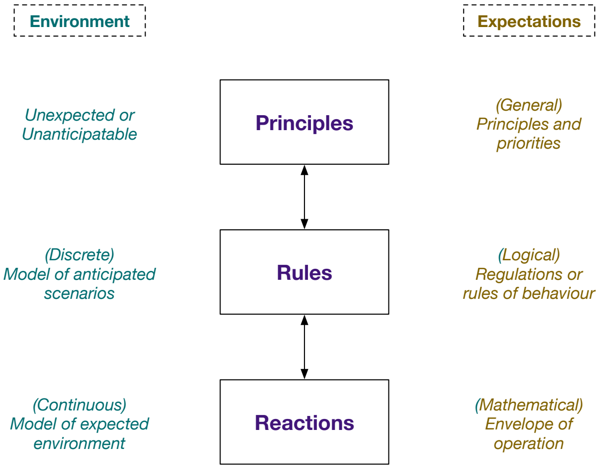
Environment
Expectations
Unexpected or Unanticipatable
Principles
(General) Principles and priorities
(Discrete) Model of anticipated scenarios
Rules
(Logical) Regulations or rules of behaviour
(Continuous) Model of expected environment
Reactions
(Mathematical) Envelope of operation

Figure 1: Three-layer autonomy framework (from [2]).

intentions/goals in autonomous systems is uniformly absent in existing standards/regulations.

3 FRAMEWORK

In order to provide a way forward, we bring together architectural/engineering issues, requirements/specification issues, and verification and validation issues, in a coherent structure presenting a reference three-level framework for autonomy in Fig. 1.

Our three-layer autonomy framework consists of: a *Reactions Layer* — comprising low-level adaptive/reactive control/response aspects; a *Rules Layer* — comprising symbolically-represented descriptions of prescribed behaviours; and a *Principles Layer* — comprising high-level, abstract, principles and priorities (again symbolic).

Together, this separates out reactive ('unconscious') interactions with the environment, from rule-compliant decisions in 'normal' situations, from ethical decisions required in unexpected and critical scenarios. The important separation here is between normal operations during which rules are followed (Rules Layer), and (unexpected/unusual) situations where our autonomous agent needs to reason about whether to violate rules, rules, for example using ethical reasoning (Principles Layer). The higher-level layers (Rules and, especially, Principles) fit well within the field of symbolic agents, particularly BDI-agents. The interactions between agent beliefs, goals, and capabilities will be crucial here.

Finally, since verification will be a cornerstone of standards and routes to certification, we discuss the range of verification (especially formal verification) options available, highlighting their advantages and drawbacks.

4 DERIVING REQUIREMENTS

One of the key challenges to certifying autonomous systems is capturing exactly how we want our systems to behave. If we do not know what is expected of the system, then how can we verify it? Unfortunately, current standards tend not to be in a form that is amenable for formalisation, since they are oriented for human use, providing declarative statements that require substantial human interpretation.

Fisher et al. [2] presents a simple process that provides guidance in identifying properties that need to be specified as verification properties for certification. The key idea is that if the autonomous system is performing tasks that are currently done by humans, then knowledge about how these humans are currently licenced can be used to help identify requirements. In doing so we consider a range of human attributes (e.g. physical capabilities, domain knowledge, regulatory knowledge), both assessed (by licencing) and assumed.

5 CASE STUDIES

In Fisher et al. [2], we provide a range of case studies to highlight how our three layer approach can be used. For example, we consider Unmanned Aerial Systems (UAS), incorporating drones, etc. Utilising our three layer framework for a UAS is straightforward.

- Autopilot **Reactions**, concerning stability, direction, etc.
- Air **Rules** to be followed.
- **Principles** for action in unanticipated/emergency situations.

As well as discussing the potential, and current regulation, concerning such systems, we address the verification possibilities and routes to certification evidence provided by our framework. For example, one approach to the potential certification of truly autonomous UAS is to show that the UAS follows the air rules that a human pilot should. Once we separate out the Rules Layer in a symbolic form, we can verify that the decisions made will be appropriate [3].

Many further case studies from across different levels of autonomy, and from very different sectors, are discussed in the article.

6 CONCLUDING REMARKS

Our work brings together hybrid control structures and the separation of different levels within a system; the identification of different requirements for control/reasoning across levels; and the verification techniques for use across such architectures. These can form the basis for analysis, design, verification and, ultimately, certification evidence for a wide range of autonomous systems.

ACKNOWLEDGMENTS

We thank the organisers and participants of the Dagstuhl 19112 seminar, http://dagstuhl.de/19112. MF was partially supported by the Royal Academy of Engineering and by EPSRC grants EP/N007565, EP/R026084, EP/R026173, EP/R026092 and EP/V026801. KYR was partially supported by NSF CAREER Award CNS-1552934, NASA ECF grant NNX16AR57G and NSF PFI:BIC grant CNS-1257011. NYS was partially supported by TAILOR, a project funded by EU Horizon 2020 research and innovation programme under grant 952215.

REFERENCES

[1] British Standards Institution (BSI). 2016. BS 8611 – Robots and Robotic Devices — Guide to the ethical design and application. https://shop.bsigroup.com/ProductDetail/?pid=000000000030320089
[2] Michael Fisher, Viviana Mascardi, Kristin Yvonne Rozier, Bernd-Holger Schlingloff, Michael Winikoff, and Neil Yorke-Smith. 2021. Towards a Framework for Certification of Reliable Autonomous Systems. *Autonomous Agents and Multi Agent Systems* 35, 1 (2021), 8. https://doi.org/10.1007/s10458-020-09487-2
[3] Matt Webster, Neil Cameron, Michael Fisher, and Mike Jump. 2014. Generating Certification Evidence for Autonomous Unmanned Aircraft Using Model Checking and Simulation. *Journal of Aerospace Information Systems* 11, 5 (2014), 258–279.