

Assessing Trustworthiness of Autonomous Systems

Greg Chance¹, Kerstin Eder¹

Trustworthy Systems Lab, University of Bristol, Bristol, UK

Abstract—As autonomous systems (AS) become more ubiquitous in society, more responsible for our safety and our interaction with them more frequent, it is essential that they are trustworthy. Assessing the trustworthiness of AS is a major challenge for the verification and development community (practitioners and researchers). Assessing trustworthiness must extend beyond conventional verification and validation (V&V) and safety-critical systems assurance, and now consider the manner in which people will interface and interact with AS across the broad range of current and future artificial intelligence (AI) applications. The meta-expression ‘trustworthiness’ is examined in the context of AS, capturing and condensing the current understanding in the literature. A list of challenges are presented in the form of a process that can be used as a trustworthiness assessment framework for AS.

1 Introduction

Autonomous systems (AS) are pervasive in current society and set to become even more so with current technological growth trends and adoption rates. Systems with embedded artificial intelligence (AI) and machine learning (ML) algorithms can be found in numerous applications from mobile phones [8], insurance pricing [7] and vacuum cleaners [9] to medical diagnostics [5], detecting structural damage to buildings [1] and predicting the shape of protein molecules [2] to name a few. There is also growing use of machine learning in a range of safety-critical systems (SCSs), for example in the aerospace and automotive industry where low reliability of these systems could result in catastrophic failure and potentially loss of life or damage to property and the environment. These safety-critical autonomous systems (SCAS) present a complex but essential challenge to the safety assurance and verification community. Conventional V&V is principally concerned with assessing the system against a set of requirements, providing guarantees of functionality and assurance of safety. But if autonomous systems are to be fully accepted into society, there must be acknowledgement of and evidence to show compliance with a broad range of *trustworthiness qualities*. This paper focuses on a reviewing what trustworthiness means for AS, how AS

can be verified as trustworthy, and the challenges associated with what that verification process may look like.

But even this may not be enough for the diverse operational domains of some AS, e.g. automated vehicles in high-density urban areas, and thinking should move beyond verification at only the system design stage, to a more continuous operational evaluation, e.g. runtime verification [ref]. Runtime verification brings other currently unresolved issues, such as suitable oracle design [ref], but some authors propose valid solutions to this using edge computing and *just in time* verification [peter thales, ref?]. Crowdsourcing verification test cases through scientific games may also be an interesting route to explore, demonstrated for an automated vehicle controller in a driving simulator [ref test gen game]

Further to this issue, are the lack of *standards* against which some trustworthy qualities should be appraised and the *methods* by which they should be evaluated. For example, there are standards for correct road driving conduct [ref UKHC] but no ethical standards by which those driving decisions should be made. Although headway is being made into developing standards for non-functional properties, such as guidelines for ethical AI [ref EU AI high level expert group] and checklists for HRI best practice [DE HRI checklist], there are still areas that need attention, such as standards and specifications for transparency and explainability, aesthetics and fairness [dhaminda22]. Additionally, there is more that can be done at the design stage to improve *verifiability*. Evidence for system correctness is essential, but this must be supported with decision explanation [koopman], whilst maintaining IPR around sensitive hardware and software algorithms [ref].

To present a defensible safety argument for AS and SCAS... trust stakeholders *trust qualities* application specific What human factors are considered eg. ISO29119

In addition to assessing the AS trustworthiness, there must also be consideration to gain, calibrate and maintain user trust in the system [4, 3], else failures related to overtrust and undertrust are possible.

In the following, related work is reviewed in Section

2 Related Work

3 Assessment Framework Vision

3.1 Assessment Methods & Corroborative Evidence

Gaining reliability assurance of SCASs using testing alone is unfeasible given the often high-dimensional operational state space. Multiple testing methodologies should be

Statements about authorship contribution. Greg Chance (e-mail: greg.chance@bristol.ac.uk), and Kerstin Eder (e-mail: kerstin.eder@bristol.ac.uk) are with the Trustworthy Systems Lab, Department of Computer Science, University of Bristol, Merchant Ventures Building, Woodland Road, Bristol, BS8 1UQ, United Kingdom.

employed where appropriate, e.g. verification, falsification and testing, [Harper Corroborative 2022] combining mutually consistent evidence from multiple and diverse assessment methods will raise the confidence in system trustworthiness.

Knowledge of the internal state of the system is often hidden, e.g. blackbox, due to IP and commercial sensitivity, but whitebox access will be essential for certain aspects of trustworthiness assessment. This may not need to reveal sensitive algorithms but just enough information through observability points in the software architecture could go a long way to understanding if automated decisions are made for the right reason[6].

4 Conclusion

References

- [1] O. Avci et al. “A review of vibration-based damage detection in civil structures: From traditional methods to Machine Learning and Deep Learning applications”. In: *Mechanical systems and signal processing* 147 (2021), p. 107077.
- [2] E. Callaway. “The entire protein universe: AI predicts shape of nearly every known protein”. In: *Nature News* 608 (2022). Accessed: 2022-08-02, pp. 15–16.
- [3] E. K. Chiou and J. D. Lee. “Trusting Automation: Designing for Responsivity and Resilience”. In: *Human Factors* (2021).
- [4] B. C. Kok and H. Soh. “Trust in robots: Challenges and opportunities”. In: *Current Robotics Reports* 1.4 (2020), pp. 297–309.
- [5] I. Kononenko. “Machine learning for medical diagnosis: history, state of the art and perspective”. In: *Artificial Intelligence in medicine* 23.1 (2001), pp. 89–109.
- [6] P. Koopman and M. Wagner. “Toward a framework for highly automated vehicle safety validation”. In: *SAE Technical Paper, Tech. Rep* (2018).
- [7] K. Kuo and D. Lupton. “Towards explainability of machine learning models in insurance pricing”. In: *arXiv preprint arXiv:2003.10674* (2020).
- [8] Medium. *Artificial Intelligence in Mobile Phones*. <https://medium.com/gobeyond-ai/artificial-intelligence-ai-in-mobile-phones-is-it-a-good-thing-fe044f20ea6c>. Accessed: 2022-08-02.
- [9] TensorFlow Blog. *Ecovacs Robotics: the AI robotic vacuum cleaner powered by TensorFlow*. <https://blog.tensorflow.org/2020/01/ecovacs-robotics-ai-robotic-vacuum.html>. Accessed: 2022-08-02.