Assessing Trustworthiness of Autonomous Systems

Greg Chance[1], Kerstin Eder[1]

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 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 [19], insurance pricing [16] and vacuum cleaners [21] to medical diagnostics [13], detecting structural damage to buildings [3] and predicting the shape of protein molecules [4] to name a few. For successful adoption of these systems then there needs to be demonstrable assurance of their safe operation which becomes increasingly difficult to in complex and changing environments. 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.

A widely held tenet is that there can never be a suitable amount of verification for complex, autonomous systems that gives complete safety-critical assurance [butler nasa]. Corroborative V&V [kerstin, harper] attempts to improve confidence through combining mutually consistent evidence from multiple and diverse assessment methods, e.g. formal, simulation, falsification, physical testing. 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 suitable oracle design [ref], but some authors propose valid solutions to this using edge computing and *just in time* verification [18, 7]. The use of *serious games* can be another interesting opportunity for building trustworthiness in complex autonomous systems and has been used in the context of mission planning for NASA [2] and also demonstrated for an automated vehicle controller in a driving simulator [ref test gen game, need to make github page, add code and youtube videos].

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 [22] 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 [15], there are still areas that need attention, such as standards and specifications for transparency [24] and explainability, asethetics and fairness [1]. Additionally, there is more that can be done at the design stage to improve *verifiability* [add ref]. Evidence for system correctness is essential, but this must be supported with decision explanation [14]. , 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 [12, 5], else failures related to overtrust and undertrust are possible.

In the following, related work is reviewed in Section

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.

1.1 Trustworthiness Qualities

Trust can be expressed in a number of ways and directions; trust the user has in the system, the objective trustworthiness of the system and the context in which the interaction between the two takes place [10]. In this research we consider only the trustworthiness of the system, but we acknowledge the importance of the other areas in which there has been much contribution from the psychology and human factors community [8, 17, 12, 5, 11]Trustworthiness of autonomous systems in this context, is the result of objective assessment of the system compared against a set of appropriate standards.

Devitt argues that reliability and accuracy are the central pilars of trustworthiness and that other properties stem from these, e.g. adaptability is a higher-order reliability [6].

2 Related Work

Floridi suggests the need for agreed upon metrics for trustworthiness of AI systems and suggests an AI Trust comparison index, metrics are needed for benchmarking AI suitability to the public.

Rudas and Haidegger also supports the idea of agreed upon metrics from the verification community that can be used to ensure reliability of complex autonomous systems [20]. Wang et al. go further and propose a theoretical framework of *tripartite trustworthiness* covering; *to-be trust* (trustfulness of an entity or structure), *to-do trust* (trust in an action or behaviour) and *system trust* (a statistical runtime evaluation of performance) and set out 18 formal definitions [23].

3 Assessment Framework Vision

3.1 Existing Standards for AS

Existing standards on verification (from Rudas 2020): P1872.1, P2817, P7000 and P7007.

Safety of autonomous systems (from Hawkins 2022): UL4000 [Underwriters Laboratories. Standard for evaluation of autonomous products, 2020] or SCSC-153B [Safety of Autonomous SystemsWorking Group. Safety assurance objectives for autonomous systems, 2022. URL: https://scsc.uk/scsc-153B]

Ethical framework for AI [9] "offer 20 concrete recommendations to assess, to develop, to incentivise, and to support good AI"

Porter2022 presents an ethical assurance argument for AS, extending the assurance case considered for safety to include ethical standards

devitt2018trustworthiness

Dhaminda to contribute here?

3.2 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 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 [14].

4 Conclusion

References

[1] D. B. Abeywickrama et al. "On Specifying for Trustworthiness". In: (2022). arXiv: 2206.11421. URL: http://arxiv.org/abs/2206.11421.

[2] B. D. Allen. "Serious gaming for building a basis of certification via trust and trustworthiness of autonomous systems". In: *2018 Aviation Technology, Integration, and Operations Conference* (2018), pp. 1–6.

[3] 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.

[4] 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.

[5] E. K. Chiou and J. D. Lee. "Trusting Automation: Designing for Responsivity and Resilience". In: *Human Factors* (2021).

[6] S. Devitt. "Trustworthiness of autonomous systems". In: *Foundations of trusted autonomy (Studies in Systems, Decision and Control, Volume 117)* (2018), pp. 161–184.

[7] K. Eder. "CyRes: towards operational cyber resilience". In: *Proceedings of the 1st International Workshop on Verification of Autonomous & Robotic Systems*. 2021, pp. 1–3.

[8] L. Floridi. "Establishing the rules for building trustworthy AI". In: *Nature Machine Intelligence* 1.6 (2019), pp. 261–262. URL: http://dx.doi.org/10.1038/s42256-019-0055-y.

[9] L. Floridi et al. "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations". In: *Minds and Machines* 28.4 (2018), pp. 689–707. URL: https://doi.org/10.1007/s11023-018-9482-5.

[10] P. A. Hancock et al. "Evolving Trust in Robots: Specification Through Sequential and Comparative Meta-Analyses". In: *Human Factors* 63.7 (2021), pp. 1196–1229.

[11] S. C. Kohn et al. “Measurement of Trust in Automation: A Narrative Review and Reference Guide”. In: *Frontiers in Psychology* 12.October (2021).

[12] B. C. Kok and H. Soh. “Trust in robots: Challenges and opportunities”. In: *Current Robotics Reports* 1.4 (2020), pp. 297–309.

[13] 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.

[14] P. Koopman and M. Wagner. “Toward a framework for highly automated vehicle safety validation”. In: *SAE Technical Paper, Tech. Rep* (2018).

[15] J. Kraus et al. “The trustworthy and acceptable HRI checklist (TA-HRI): questions and design recommendations to support a trust-worthy and acceptable design of human-robot interaction”. In: *Gruppe. Interaktion. Organisation. Zeitschrift für Angewandte Organisationspsychologie (GIO)* (2022), pp. 1–21.

[16] K. Kuo and D. Lupton. “Towards explainability of machine learning models in insurance pricing”. In: *arXiv preprint arXiv:2003.10674* (2020).

[17] J. D. Lee and K. A. See. “Trust in automation: Designing for appropriate reliance”. In: *Human Factors* 46.1 (2004), pp. 50–80.

[18] C. Maple et al. *CyRes – Avoiding Catastrophic Failure in Connected and Autonomous Vehicles (Extended Abstract)*. 2020. URL: https://arxiv.org/abs/2006.14890.

[19] 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.

[20] I. Rudas and T. Haidegger. “Verification, trustworthiness and accountability of human-driven autonomous systems”. In: *2021 IEEE International Conference on Autonomous Systems (ICAS)*. IEEE. 2021, pp. 1–1.

[21] 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.

[22] UK Driving Standards Agency. *The Official Highway Code*. Her Majestys Stationery Office, 2012.

[23] Y. Wang et al. “A Tripartite Theory of Trustworthiness for Autonomous Systems”. In: *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics* 2020-October (2020), pp. 3375–3380.

[24] A. F. Winfield et al. “IEEE P7001: A proposed standard on transparency”. In: *Frontiers in Robotics and AI* (2021), p. 225.