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 mandatory challenge for the verification and development community. This will require appropriate standards and suitable metrics that may serve to objectively and comparatively judge trustworthiness of AS across the broad range of current and future applications. The meta-expression 'trustworthiness' is examined in the context of AS capturing the relevant qualities that comprise this term 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 [30], insurance pricing [26] and self-driving vehicles to medical diagnostics [22], detecting structural damage to buildings [4] and predicting the shape of protein molecules [7] to name a few. For successful adoption of these systems then there needs to be demonstrable assurance of their trustworthy operation which becomes increasingly difficult 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 present a complex but essential challenge to the safety assurance and verification community.

Verification and validation (V&V) is the process to gain confidence in the correctness of a system relative to its requirements. Prior to, and separate from verification, a specification must clearly define the trustworthy operational behaviour of the system and many challenges are associated with this task for autonomous systems [1]. 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*. A major challenge in verifying this broad category of trustworthiness qualities, are the lack of standards and regulations against which they can be evaluated. And whereas verification methods of assessing, for example, functional correctness are relatively mature, there also exists the challenge of developing robust assessment methodologies and metrics for these more nuanced trustworthiness qualities. This paper focuses on reviewing what trustworthiness means in the fields of AS, robotics, HRI and Cyber-Physical Systems (CPS). An assessment process is presented along with the challenges associated with practical application.

1.1 V&V for AS in complex environments

A widely held tenet is that there can never be a suitable amount of verification that gives complete assurance for complex, safety-critical autonomous systems, although limits on reliability rates have been proposed [6]. Corroborative V&V [37] attempts to improve confidence through combining mutually consistent evidence from multiple and diverse assessment methods, e.g. formal, 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 such as runtime verification. Runtime verification brings other currently unresolved issues, such suitable oracle design [28], but some authors propose valid ideas to this using edge computing and cloud-based verification [29, 11]. 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 in a self-driving vehicle controller leveraging the power of crowdsourcing for test generation [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 [35] but no ethical standards by which those driving decisions should be made [ref like trolley problem?]. 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], checklists for HRI best practice [24] and transparency [38], there are still areas that need attention, such as standards for adaptability, cooperation and fairness [1]. Where standards are lacking or immature will require engagement

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.

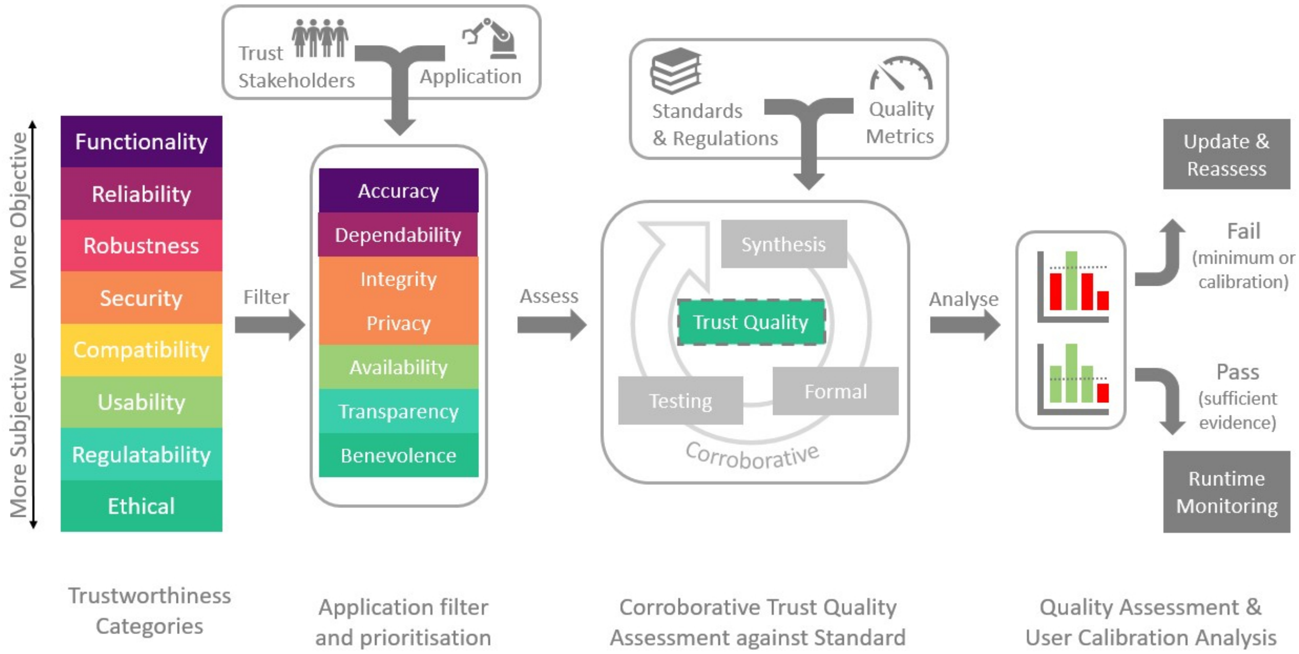

Fig. 1: AS trustworthiness assessment process

with *trust stakeholders*, expert steering groups that can define and prioritise the necessary trustworthiness qualities for each subject domain or application.

Additionally, there is more that can be done at the design stage to improve *verifiability* [add ref]. Evidence for functional correctness is essential, but this must be supported with decision explanation [23] whilst maintaining intellectual property rights around, for example, sensitive software algorithms and trade secrets [ref].

In addition to assessing the AS trustworthiness, there must also be consideration to gain, calibrate and maintain user trust in the system [21, 8], as miscalibration of trust between system and user can have serious consequences [21].

1.2 Trustworthiness Qualities

As computing and automation has developed, systems are now both more capable and users more reliant on them. This extension of capability has resulted in a broadening of the terms which encompass trustworthiness, as, for example, the important trustworthiness qualities of a calculator may be less numerous than those of a medical decision support system. Advancement in automation then, has led us to question and challenge these new capabilities, or, with more automation comes more responsibility [40].

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 [15]. Trustworthiness can also been described as a the probability that a system holds some established property or quality, and that greater trustworthiness begets greater likelihood that the system may exhibit that quality. In this research we consider the trustworthiness of the system and the specific qualities that must be demonstrated, but we acknowledge the importance of the other mechanisms where human-system trust can be gained or lost in which there has been much contribution from the HRI, psychology and human factors community [12, 27, 21, 8, 20, 24]. Trustworthiness of autonomous systems in the context of this work then, results from objective assessment of the system with respect to a set of appropriate standards. There has been much academic deliberation on the specific qualities that comprise trustworthiness of AS, specifically for AI [33, 39] and HRI [24, 3] Devitt argues that reliability and accuracy are the two central pillars of trustworthiness of AS and that all other properties stem from these, for example, stating that adaptability and redundancy are higher-order properties of reliability [10].

[34] state 5 facets of trustworthy software: Safety: The ability of the software to operate without causing harm to anything or anyone.Reliability: The ability of the software to operate correctly. Availability: The ability of the software to operate when required. Resilience: The ability of the software to recover from errors quickly and completely. Security: The ability of the software to remain protected against the hazards posed by malware, hackers or accidental misuse.

There is, in fact, a spectrum of qualities that comprise user trust in an autonomous system...(leads into categories)

A summary of AS trustworthy qualities in the literature can be found at [https://github.com/TSL-UOB/TAS-Verif]

1.3 Ontology of AS Trustworthiness Qualities

An ontology of trustworthiness qualities is presented. This is an independent set of characteristics, where one category is not necessary influenced or related to its neighbours, that supports clarity of the assessment [connett 2018]

(Lee & See 2004) performance, process, purpose, which although are broadly capturing of all trustworthy qualities miss some nuance to give a complete and expressive account of trustworthiness.

avizienis2004basic a set of general concepts are required for dependable and secure computing, which may cover a wide range of applications and system failures, which comprise; availability (readiness for correct service), reliability (continuity of correct service), safety (absence of catastrophic consequences on the user(s) and the environment), integrity (absence of improper system alterations) and maintainability (ability to undergo modifications and repairs).

Thiebes et al. argue for five foundational principles of trustworthy AS: beneficence (doing good), non-maleficence (not harming), autonomy (preserving human decision making), justice (fair and reasonable), and explicability (easily understood) [33]. These are based on and related to other discussion on ethically principled foundations of trustworthiness, Floridi 2018 [13], and should be weighted appropriate to the strong international contribution to this area Jobin 2019 [18] but fail to capture the full trustworthiness spectrum of qualities presented here.

2 Assessment Framework Vision

Below is a checklist for assessemnt of AS trustworthiness: **Metrics** 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 [32]. 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 [36]. Garbuk presents the idea of *applied intellimetry* to assess the quality of AI systems by formulating a list quality characteristics in a functional characteristic vector [14]. Kaur et al. suggest explainability metrics based on the euclidean distance between the system output compared to a panel of experts [19]. trustworthiness of computer systems using metrics designed to assess security, trust, resilience and agility [9]

Bolster and Marshall proposes the idea of *multi-vector trust metrics* for networks of autonomous systems, indicating that the use of *grey relational analysis*, a theory to describe and model uncertainty, could be beneficial for combining temporally sparse, low fidelity metrics with unknown statistical distributions [5].
Verification Methods: Kress-Gazit et al. state that assessment in the correctness of AS can be broken down into four approaches: synthesis of correct-by-construction systems, formal verification at design time, runtime verification or monitoring, and test-based methods [25].

2.1 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 [13] "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

What human factors are considered eg. ISO29119

Dhaminda to contribute here?

2.2 Future Challenges in Standards

Riaz et al. [31] suggest the idea of using social norms and human emotions as a standard by which better self-driving controllers may be developed. This idea sets the way for not just development of higher functioning AS, but also standards of trustworthiness by which they can be judged. Although there is much scholarly work on the theory and modelling of social norms, e.g. [17], there is yet to be published a standard that could be used to objectively assess an autonomous system.

In some cases, e.g. driving, legislation on appropriate conduct is presented to society in the form of guidelines such as the UKHC in the UK [35] but must be translated to a computer readable format to act as an appropriate standard, or set of assertions [16], if these guidelines can be used to assess AS trustworthiness. A similar process will have to be undertaken for other standards which have yet to be defined, e.g. cooperation, fairness or verifiability, to ensure all aspects of trustworthiness can be assessed.

2.3 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 [23].

Table 1: Trustworthiness qualities ontology

Trustworthiness Quality Category	Definition
Functionality	Ability of a system to not enter a failure mode, to be able to execute tasks required of it without fault, to achieve a goal state (liveness), and do so within permitted use of resources.
Compatibility	Degree to which the system can exchange information with users or other systems, be transferred to other environments and the ability to share the same environment with other autonomous agents.
Usability	Extent to which the system is available and responsive and can be used to achieve specified goals with effectiveness, and satisfaction in a specified context.
Reliability	Degree to which the system performs specified functions under specified conditions for a specified period of time in a consistent manner.
Security	Protection against intentional subversion or forced failure, malicious access, use, modification, destruction, or disclosure. Defining, achieving, and maintaining confidentiality, integrity, availability, non-repudiation, accountability, authenticity, and reliability of the system.
Robustness	Ease with which the system can overcome adverse conditions and be maintained or modified to change or add capabilities or to operate at new scales, correct faults or defects, improve performance or other attributes and to adapt to new environments.
Ethical	Ability to demonstrate beneficence and non-maleficence, fair and reasonable behaviour, to preserve human decision making and be easily understood
Regulatability	Ease in which the system is verifiable, readable, explainable, transparent and understandable in a manner to support regulation, appropriate trustworthy metrics and specifications.

3 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] D. Atkinson et al. "Trust in computers and robots: The uses and boundaries of the analogy to interpersonal trust". In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting.* Vol. 56. 1. SAGE Publications Sage CA: Los Angeles, CA. 2012, pp. 303–307.

[4] 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.

[5] A. B. Bolster and A. Marshall. "A multi-vector trust framework for autonomous systems". In: *AAAI Spring Symposium - Technical Report* SS-14-04.April 2014 (2014), pp. 17–19.

[6] R. W. Butler and G. B. Finelli. "The Infeasibility of Quantifying the Reliability of Life-Critical Real-Time Software". In: *IEEE Transactions on Software Engineering* 19.1 (1993), pp. 3–12.

[7] 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.

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

[9] J.-H. Cho et al. "Stram: Measuring the trustworthiness of computer-based systems". In: *ACM Computing Surveys (CSUR)* 51.6 (2019), pp. 1–47.

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

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

[12] 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.

[13] 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.

[14] S. V. Garbuk. "Intellimetry as a way to ensure AI trustworthiness". In: *2018 International Conference on Artificial Intelligence Applications and Innovations (IC-AIAI).* IEEE. 2018, pp. 27–30.

[15] 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.

[16] C. Harper et al. "Safety Validation of Autonomous Vehicles using Assertion-based Oracles". In: *arXiv preprint arXiv:2111.04611* (2021).

[17] M. Hechter and K.-D. Opp. *Social norms.* Russell Sage Foundation, 2001.

[18] A. Jobin, M. Ienca, and E. Vayena. "The global landscape of AI ethics guidelines". In: *Nature Machine Intelligence* 1.9 (2019), pp. 389–399.

[19] D. Kaur et al. "Trustworthy explainability acceptance: A new metric to measure the trustworthiness of interpretable AI medical diagnostic systems". In: *Conference on Complex, Intelligent, and Software Intensive Systems.* Springer. 2021, pp. 35–46.

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

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

[22] 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.

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

[24] 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.

[25] H. Kress-Gazit et al. "Formalizing and guaranteeing human-robot interaction". In: *Communications of the ACM* 64.9 (2021), pp. 78–84.

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

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

[28] M. Leucker and C. Schallhart. "A brief account of runtime verification". In: *Journal of Logic and Algebraic Programming* 78.5 (2009), pp. 293–303. URL: http://dx.doi.org/10.1016/j.jlap.2008.08.004.

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

[30] 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.

[31] F. Riaz et al. "A collision avoidance scheme for autonomous vehicles inspired by human social norms". In: *Computers and Electrical Engineering* 69 (2018), pp. 690–704.

[32] 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.

[33] S. Thiebes, S. Lins, and A. Sunyaev. "Trustworthy artificial intelligence". In: *Electronic Markets* 31.2 (2021), pp. 447–464.

[34] Trustworthy Software Foundation. *TS Framework*. http://www.tsfdn.org/ts-framework/. Accessed: 2022-08-17.

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

[36] 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.

[37] M. Webster et al. "A corroborative approach to verification and validation of human–robot teams". In: *The International Journal of Robotics Research* 39.1 (2020), pp. 73–99.

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

[39] J. M. Wing. "Trustworthy AI". In: *Communications of the ACM* 64.10 (2021), pp. 64–71.

[40] V. Yazdanpanah et al. "Responsibility research for trustworthy autonomous systems". In: *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS* 1 (2021), pp. 57–62.