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 [**medium˙ai˙phones**], insurance pricing [**kuo2020towards**] and vacuum cleaners [**tf˙vacuum**] to medical diagnostics [**kononenko2001machine**], detecting structural damage to buildings [**avci2021review**] and predicting the shape of protein molecules [**alpha˙fold**] 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.

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.

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 [**CyRes20**, **eder2021cyres**]. 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 [**Allen2018**] 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 [**highwayCode**] 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 [**kraus2022trustworthy**], there are still areas that need attention, such as standards and specifications for transparency [**winfield2021ieee**] and explainability, asethetics and fairness [**Abeywickrama2022**]. 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 [**koopman2018toward**]. , whilst maintaining IPR around sensitive hardware and software algorithms [ref].

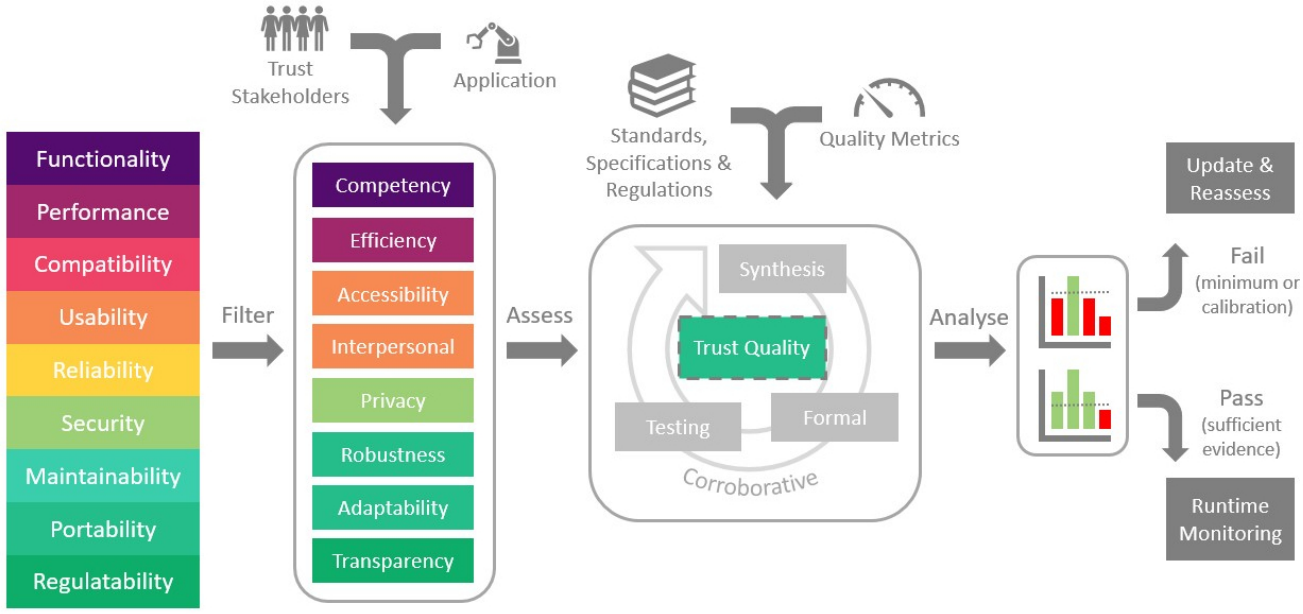

Fig. 1: AS trustworthiness assessment process

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 [**kok2020trust**, **Chiou2021**], else failures related to overtrust and undertrust are possible.

In the following, related work is reviewed in Section

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 [**Hancock2021**]. 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 [**Floridi2019**, **Lee2004**, **kok2020trust**, **Chiou2021**, **Kohn2021**, **kraus2022trustworthy**]. 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 [**Thiebes2021**, **Wing2021**] and HRI [**kraus2022trustworthy**, **atkinson2012trust**] Devitt argues that reliability and accuracy are the central pillars of trustworthiness and that other properties stem from these, stating that adaptability and redundancy are higher-order properties of reliability [**devitt2018trustworthiness**]. Thiebes et al. argue for five foundational principles of trustworthy AS (1) beneficence, (2) non-maleficence, (3) autonomy, (4) justice, and (5) explicability [**Thiebes2021**].

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 [**Rudas2020**]. 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 [**Wang2020**].

3 Assessment Framework Vision

Below is a checklist for assessemnt of AS trustworthiness: **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 [**kress2021formalizing**].

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 [**Floridi2018**] "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 Future Challenges in Standards

Riaz et al. [**Riaz2018**] 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. [**hechter2001social**], 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 [**highwayCode**] but must be translated to a computer readable format to act as an appropriate standard, or set of assertions [**harper2021safety**], 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.

3.3 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 [**koopman2018toward**].

4 Conclusion