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 [**mediumaiphones**], insurance pricing [**kuo2020towards**], vacuum cleaners [**tf˙vacuum**] and self-driving vehicles 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 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 [**Abeywickrama2022**]. 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.

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 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 [**Butler1993**]. Corroborative V&V [**webster2020corroborative**] 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 [**Leucker2009**], but some authors propose valid ideas to this using edge computing and cloud-based 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 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 [**highwayCode**] 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 [**kraus2022trustworthy**] and transparency [**winfield2021ieee**], there are still areas that need attention, such as standards for adaptability, cooperation and fairness [**Abeywickrama2022**]. Where standards are lacking or immature will require engagement 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 [**koopman2018toward**] 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 [**kok2020trust**, **Chiou2021**], as miscalibration of trust between system and user can have serious consequences [**kok2020trust**].

2 Trustworthiness of AS Qualities (TASQ)

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, as some commentary has noted: with more automation comes more responsibility [**Yazdanpanah2021**].

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**]. 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 [**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 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 [**devitt2018trustworthiness**].

[**ts'foundation**] 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.

2.1 Ontology of AS Trustworthiness Qualities

A trust ontology can be a useful definition to identify an independent set of important quality characteristics, where one category is not necessary influenced or related to its neighbours. These categories can be used to support clarity of communication and understanding of issues pertaining to, and of judgement in the assessment of, trustworthiness of autonomous systems.

Lee & Moray propose the categories for trust in automation: performance (consistent and stable behaviour), process (qualities or characteristics that govern behaviour), purpose (underlying motive or intent) and foundation (fundamental assumptions of natural and social order). These categories broadly capture the full gamut of trustworthy qualities but may be too broad and abstract for practical assessment purposes.

Avizienis et al. proposes that a set of general concepts are required for dependable and secure computing, which may cover a wide range of applications and system failures, comprising; availability (readiness for correct service), reliability (continuity of correct service), safety (absence of catastrophic consequences on the user and the environment), integrity (absence of improper system alterations) and maintainability (ability to undergo modifications and repairs) [**avizienis2004basic**]. The focus here is on functionality and usability, but these categories may be too specific to computing and neglect verifiability and ethical considerations around AI.

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) [**Thiebes2021**]. These are based on and related to numerous other discussion on ethically principled foundations of trustworthiness and there is evidence of strong international collaboration and motivation in this area [**Floridi2018**, **jobin2019global**]. Whilst these categories are very important and capture ethical and regulatory considerations, they fail to capture the aspects of functionality and dependability of other voices in the community.

Cho et al. propose a STRAM ontology for measuring the trustworthiness of computer systems, based around

four sub-metrics of: security (availability, confidentiality, integrity), trust (predictability, safety, reliability), resilience (adaptability, fault-tolerance, recoverability) and agility (efficiency, usability, timeliness), although again functional aspects are missing with the main focus being on security.

2.1.1 Ontologies within Existing Standards

The international standard ISO/IEC/IEEE 29119 describes software and systems engineering and part-4 covers software testing techniques and outlines 8 areas that testing should focus around: Functional Stability, Performance Efficiency, Compatibility Usability, Reliability, Security, Maintainability, Portability [**ISO29119**]. This standard is primarily focused on software testing and so some of these categories, although useful, includes jargon specific to computing systems which were not repeated in any other literature pertaining to AS more generally and therefore less user-friendly. However, part-13 of ISO29119 sets out standards specifically for testing AI-based software systems which extends these quality characteristics to include AI-specific qualities such as: flexibility (range of behaviours), adaptability (ease of modification or achieving flexibility), autonomy (unsupervised ability and level of control), evolution (behaviour adaptation over time), bias (e.g. due to discrimination, historic bias, uneven sampling), transparency (access to data and algorithms and decision interpretability) and determinism (same output for given input) as well as consideration to ethical specifications and side-effects such as reward hacking.

FUTURE:ISO standard on "quality model for AI-based systems"

DIN SPEC 92001 is a standard to help ensure quality in AI systems. Part 2 of the standard (92001-2) describes three pillars responsible for AI quality, namely: functionality and performance, robustness and comprehensibility - look into

The NIST Framework for Cyber-Physical Systems [ref] details a list of trustworthy 'aspects' and 'concerns' in addition to operational and business concerns for CPS and also includes some excellent case studies to show a complete end-to-end analysis, whilst Balduccini goes on to draw reasoning about the trustworthy properties set out in the framework in a UML/XML language [ref].

Dhaminda to contribute here?

A spectrum of qualities is presented that captures the broad definition of trustworthiness of AS from the literature, see Table 1. A full list of the quality terms reviewed can be found at [**tsl˙git**].

3 Application Criticality

What is not considered a great deal in the literature is application criticality; what application the AS is used for and if this should change the significance of specific trustworthiness qualities. Applications will need more emphasis on certain trustworthy qualities depending on where the system is most vulnerable to violating trust. Arianna, is there a focus in ethics on where the power is held? and whom can be prejudiced/discriminated against? For example, a self-driving vehicle, or indeed any safety critical system, must have emphasis on safety, possibly to the detriment of other qualities.

DIN SPEC 92001-1 describes a *quality meta model* and distinguishes between high risk systems that have safety, security, privacy and ethical relevance and those that do not (low risk), delineating applications into two risk classes [**Englisch2019**] which can be assessed using an appropriate risk assessment process, e.g. HAZOP, FMEA, SHARD. High risk applications must commit evidence of system trustworthiness based on these categories (or must be justified) whilst low risk systems are less strict.

Whilst this DIN SPEC approach is commendable, it does not go far enough to filter trust qualities based on the application and identify the key qualities required for assessment. The risk assessment process can be used to identify those qualities which are most pertinent to the application which can be prioritised, included or discarded entirely. For example, there may be an application with strong safety requirements but little to none regarding privacy.

Fisher 2021 asks if there are some rules that are more important that others, are all rules born equal? Context and application, social and cultural norms will all influence this answer. An example may be the ease with which breaking the speeding limit is observed in driving behaviour, but other driving conduct rules are broken less often, such as driving through a red traffic signal light.

3.1 Decision Making Complexity

Within application criticality comes decision making complexity, the complexity of the process the AS must navigate in order to achieve the task goal. This could be considered in terms of the system and the constraints in the environment or action space to allow or prevent the system reaching a number of potential future states. A autonomous vacuum cleaner, for example, is physically constrained to a small 2-dimensional plane (area to be cleaned) and can successfully function with a small action space (stop, rotate, drive) where decisions are reactively made based on a few simple sensor inputs (avoid close object). Contrast this to, for example, any system that requires a perception stack to interpret dynamic physical scenes, such as a self-driving vehicle, which must identify and extrapolate objects and their future states (pedestrians), environmental conditions (road furniture, fog) and static or temporary rules that require interpreting (road signs, traffic cones) which will all contribute to a decision, which is also exacerbated by the potential harm that can result if a wrong decision is made. Decision making complexity needs to be considered when judging the risk level of the application, and the associated trust qualities should be elevated accordingly, where applications that have high risk associated with particular qualities, those qualities should be elevated to higher risk levels and be assessed accordingly (see Section 5).

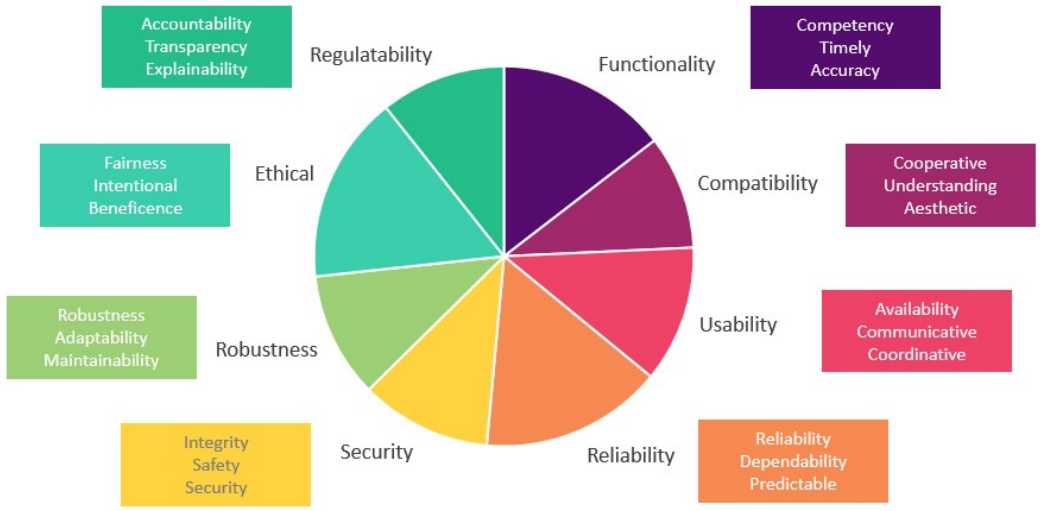

Fig. 1: Analysis of trust quality terms in the literature placed into categories, breakout box shows most cited words from each category.

3.2 Potential harm (from failure)

Qualities that are principally connected to the functionality of the AS, are required for trustworthiness. A vacuum cleaner with poor coordination can do limited harm to users, but the same lacking quality in a robotic assistant, say, may fail to be accepted as trustworthy and also cause potential harm. Therefore, the trust quality must be elevated to a high risk level. A risk assessment process should be able to identify such critical properties that can follow through the verification process.

4 Automation Level

The level of automation is an important consideration which relates to what qualities the system should present, our reliance and vulnerability to the system and hence the criticality of the application. A good description of automation levels are given by SAE International [**SAEJ3016**]. Fisher et al. describe *automation scope* which, alongside the level of automation, describes the sophistication of potential system actions and the ability to achieve complex task goals [**Fisher2021**]. Alongside scope, agency (independent acting or decision making) and whether to be reactive (responding to a situation or stimulus) or proactive (creating a situation) in action decision are also factors to consider within automation level. Greater scope, greater responsibility [ref] But automation scope is simply an aspect of non-functional AS qualities, of which compatibility, which includes coexistence and harmony, are aspects of automation that would naturally extend the scope of the system.

5 Assessment Framework Vision

Below is a checklist for assessment of AS trustworthiness:

5.1 TASQ Categories

After reviewing the literature, a comprehensive ontology has been derived based on TASQ.

5.2 Application Criticality Analysis

Which TASQ to include in your application? Are all qualities relevant? How can quality relevance be objectively assessed? What inputs are required from trust stakeholders? How does automation level affect the choice?

5.3 TASQ Assessment with Corroborative V&V

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**]. This assessment uses all four approaches both at design time, during operation (runtime) and as a consequence of development iteration or update.

5.4 Standards and Regulations

What standard for each quality? Is there a relation/pattern? Can we map standards to qualities or is it application based?

5.5 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 [**Rudas2020**]. Wang et

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, nonrepudiation, 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.

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**]. 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 [**garbuk2018intellimetry**]. Kaur et al. suggest explainability metrics based on the euclidean distance between the system output compared to a panel of experts [**kaur2021trustworthy**]. trustworthiness of computer systems using metrics designed to assess security, trust, resilience and agility [**cho2019stram**]

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 [**Bolster2014**].

For data-centric and highly objective measures of trust, operators such as accuracy, precision and recall can be useful for functionality metrics. Questions over what is the ground truth is, a sample of the real world, artificially augmented, and if the data is ethically sourced are all additional factors to be considered. We may even be more abstract and use *task completion* - this will confluence with other assessment areas.

5.6 Analysis & Visualising Trust

Whereas metrics for the development and regulation community may be highly technical, those that are public facing must be presented in a user-friendly manner that will attain and hold trust in the user, whilst maintaining the meaning of salient information without clouding in jargon [ref].

As suggested by Floridi, public confidence in AI-based systems could be bolstered with an internationally recognised index for trustworthy AI, such as a *trust comparison index* or AI *star rating*. A vision of this rating is shown in Table XXX. Much like an internet browser may convince users of their safety on a web page through some means, e.g. icons, coloured indicators, a similar approach could be taken with AS applications be that through cyber or physical means.

Corroborative, mutually consistent evidence from diverse methodologies provides assurance that is greater in quality than evidence from single sources and is the only process that can result in the highest TASQ rating. Evidence that fails to corroborate, although does not fail in itself, may be considered equivalent to single-source evidence and therefore, although not a failure, can only provide evidence for low risk applications.

If the risk assessment analysis deems the quality not to be critical then the minimum threshold for the specific metric of that quality can be relaxed for a passing grade.

5.7 Insufficient Trust: Update & Reassessment

5.8 Sufficient Trust: Monitoring

5.9 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.

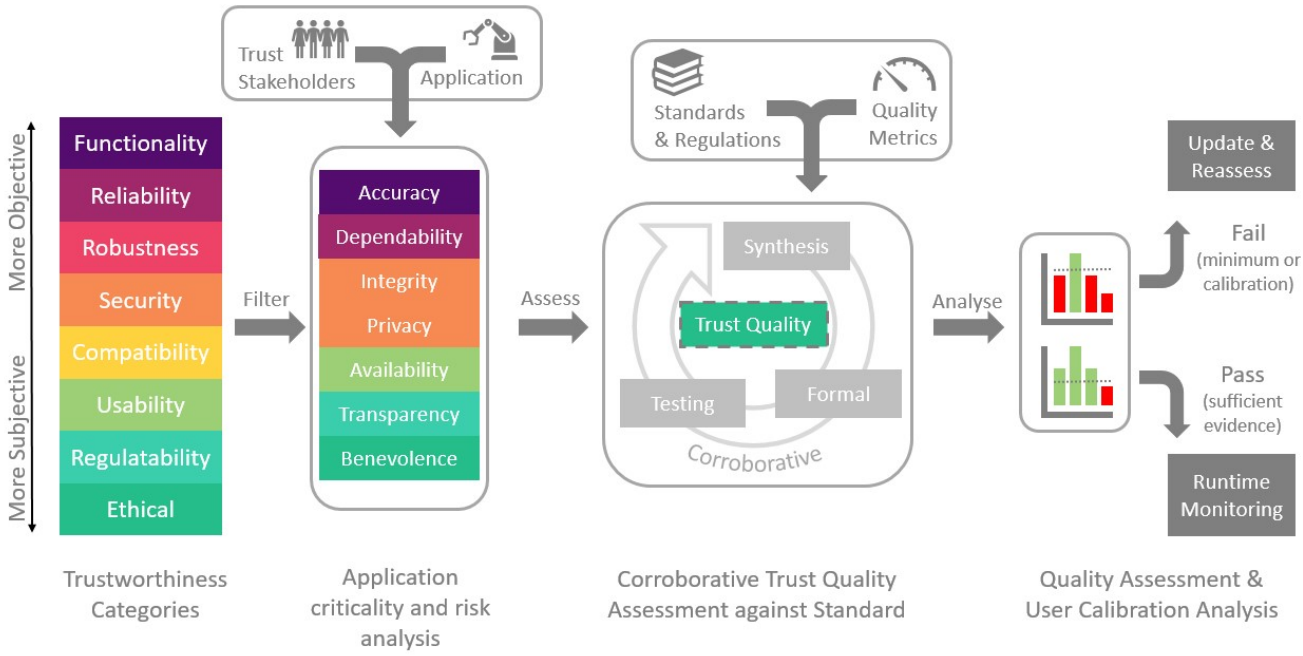

Fig. 2: AS trustworthiness assessment process

Table 2: Trustworthiness Autonomous Systems Quality (TASQ) star rating comparison index

TASQ Rating	Assessment Description	Application Class
★	Single assessment method, meets minimum compliance with standards for low risk applications	low risk applications only
★★	1-2 assessment methods where appropriate, meets recommended compliance level and attempt to calibrate user trust	low-moderate risk applications
★★★	evidence from at least 2 diverse assessment methods, meets full compliance guidelines and extensive user trust calibration	any risk level applications

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.

5.10 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**].

6 Conclusion