# Assessing Trustworthiness of Autonomous Systems

Greg Chance[1], Dhaminda Abeywickrama[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 are systems that involve software applications, machines and people, which are capable of taking actions with no or little human supervision [47]. 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 [46], insurance pricing [39], vacuum cleaners [56] and self-driving vehicles to medical diagnostics [35], detecting structural damage to buildings [5] and predicting the shape of protein molecules [10] to name a few. For successful adoption and reliance upon AS, there needs to be demonstrable assurance of their trustworthy operation which becomes increasingly difficult for complex systems with more ambitious *automation scope* and being used for applications with greater criticality.

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 trusted into society, there must be acknowledgement of, and evidence to show, compliance with a broad range of *trustworthiness qualities*. A trustworthiness quality is defined here as a non-functional system property that promotes reliance or enables adoption in that system with respect to the views of the *trust stakeholders*. A survey of the literature on trustworthy AS has identified many qualities that have been grouped into an ontology and includes properties such as robustness, security and accountability. Trust stakeholders have a vested interest in the reliable operation of the system or otherwise seek assurance of specific system properties from a variety of perspectives, such as: end users or operators, regulators, developers or system manufacturers, unintended users and, to some extent, the environment.

A major challenge in verifying this broad category of trustworthiness qualities, is the availability 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, e.g. fairness, beneficence. In addition, systems with operational adaptation or evolving functionality should have some level of runtime monitoring of appropriate metrics to ensure the system does not move to an untrustworthy state or condition.

This paper focuses on reviewing what trustworthiness means in the field of AS, including robotics, HRI and Cyber-Physical Systems (CPS), and how the application, criticality and automation scope influence trust assessment. Trustworthiness of the system must also be reciprocated with user trust in the system for successful adoption and subsequent reliance [41]. User trust may consider of other factors contributing, such as the broader epistemological or sociotechnical aspects of trust but these will not be considered here, nor the process of gaining, maintaining or preventing the erosion of the *trust agreement* between the user and the system, of which there are many excellent discussions, e.g. see [33, 41, 34, 11, 20]. We focus on technical aspects of trustworthiness relating to the system and outline the challenges associated with practical application of a presented assessment process.

This document is structured as follows: ...

### 1.1 Verification and Validation for AS

Attaining complete assurance of any complex system is challenging and, in some cases where the input parameter space is large, it may be intractable. However, there are multiple techniques and approaches that the verification engineer can use to seek assurances against certain system properties, such as formal modelling, testing, synthesis and runtime verification [38]. These techniques

*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.

and the Vmodel of verification have been used to successfully provide assurance of software functionality for several decades [17]. But as the functionality, ambition and level of automation of deployed systems grows, the assessment methodology must also adapt and account for these new responsibilities. This may include aspects such as security, privacy and other ethical concerns, highlighted by, for example, the need for better online security and protection against web search surveillance [48]. There are emerging developments in the areas of ethical AI assessment and assurance, for example Porter et al. propose a principles-based ethical argument (PBEA) framework for reasoning about the overall ethical acceptability of an AS [49].

Corroborative V&V [62] attempts to improve confidence through combining mutually consistent evidence from multiple and diverse assessment methods, e.g. formal, testing [55]. But even this may not be enough for the diverse operational domains of some AS and thinking should move beyond design time verification, to a more continuous operational evaluation such as *runtime verification*. Runtime verification brings other currently unresolved issues, such suitable oracle design [42], but some authors propose valid ideas to this using edge computing as a cloud-based verification authority [45, 15].

Further to these issues, 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 [60] but ethical standards by which those driving decisions should be made do not exist or are just emerging [8] . An interesting dilemma is if maximising the trust in qualities result in conflict, and how these conflicts can be ethically resolved. Although headway is being made into developing standards for non-functional properties, such as guidelines for ethical AI [20], checklists for HRI best practice [37] and transparency [64], 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 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 [36] 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 [34, 11], as miscalibration of trust between system and user can have serious consequences [34].

## 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 [66].

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 [22]. 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 [19, 41, 34, 11, 33, 37]. 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 [57, 65] and HRI [37, 4] 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 [13].

[58] 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) [6]. 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) [57]. 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 [20, 30]. 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.

### 2.2 Standards for Autonomous Systems and Trustworthiness Properties

The functionality of an autonomous system (i.e. what it is meant to do, what it does, and what it could do) evolves or changes over time. One of the main issues of adopting current standards and regulations with autonomous systems is the lack of consideration to the notions of *uncertainty* and *autonomy* [18]. Most conventional processes for defining system requirements assume that these are fixed and can be defined in a complete and precise manner before the system goes into operation [1]. Also, in existing standards and regulations, the notion of autonomy is not their most characterizing feature where they are neither driven nor strongly influenced by it [18]. Most existing standards are either implicitly or explicitly based on the V&V model, which moves from requirements through design onto implementation and testing before deployment [29]. However, this model is unlikely to be suitable for systems with the ability to adapt their functionality in operation; e.g. through interaction with other agents and the environment (e.g. as is the case with swarms); or through experience-driven adaptation as is the case with machine learning [1]. we have not yet mentioned systems with evolving functionality, this can be put in the introduction Autonomous systems with evolving functionality follow a different, much more iterative life-cycle. Thus, there is a need for new standards and assurance processes that extend beyond design time and allow continuous certification at runtime [53]. In this context, lately, there have been several standards and guidance introduced by several industry committees and research groups. Now we provide an overview of several key efforts with any trustworthiness properties or ontologies supported by them.

In 2016, the British Standards Institution introduced the *BS 8611* standard that provides a guide to the ethical design and application of robots and robotic systems [9]. Then, IEEE through its initiative Global Initiative on Ethics of Autonomous and Intelligent Systems initiated the development of a series of standards to address autonomy, ethical issues, transparency, data privacy and trustworthiness (IEEE P70XX, for e.g. IEEE P7001, P7007, P7010). *IEEE P7001* standard describes measurable, testable levels of transparency for autonomous systems so that they can be objectively assessed and levels of compliance determined[63]. This standard outlines five stakeholder groups, and for each group it explains the structure of the normative definitions of levels of transparency. *IEEE P7001* can be applied to assess the transparency of an existing system using a process of System Transparency Assessment, or to specify transparency requirements for a system prior to its implementation using a System Transparency Specification. Meanwhile, the goal of *IEEE P7007* standard is to assist in the ethically-driven methodologies for the design of robots and automation systems [26]. For this, it provides a set of ontologies with different abstraction levels of concepts, definitions, axioms and use cases. IEEE P7001 mentions about a transparency concern, which is a property representing an explanation topic (e.g. fairness, safety, legality, reliability, accountability, responsibility, predictability, comprehensibility, justifiability, viability, coordination) describing the reason for explanations of agent behaviours. *IEEE P7010* standard is used to measure the impact of AI or autonomous and intelligent systems on humans [27].

There are several standards and guidance related to machine learning in aeronautics, automotive, railway and industrial domains, e.g. AMLAS, EASA concept paper, DEEL white paper, AVSI report, LNE certification and UL 4600 standard [31].

*Assurance of Machine Learning for use in Autonomous Systems (AMLAS)* provides guidance on how to systematically integrate safety assurance into the development of the machine learning components based on offline supervised learning [24]. AMLAS contains six stages, and the assurance activities are performed in parallel to the development of machine learning component. The process is iterative by design and feedback is used to update previous stages. In AMLAS, the safety requirements are always based on performance and robustness of the machine learning model. In a related work to AMLAS [3], the authors identify several phases in a machine learning life cycle (data management, model learning, verification and deployment) with their associated data. From an assurance view point, they consider several key properties the models generated by learning should exhibit: performance, robustness, reusability and interpretability.

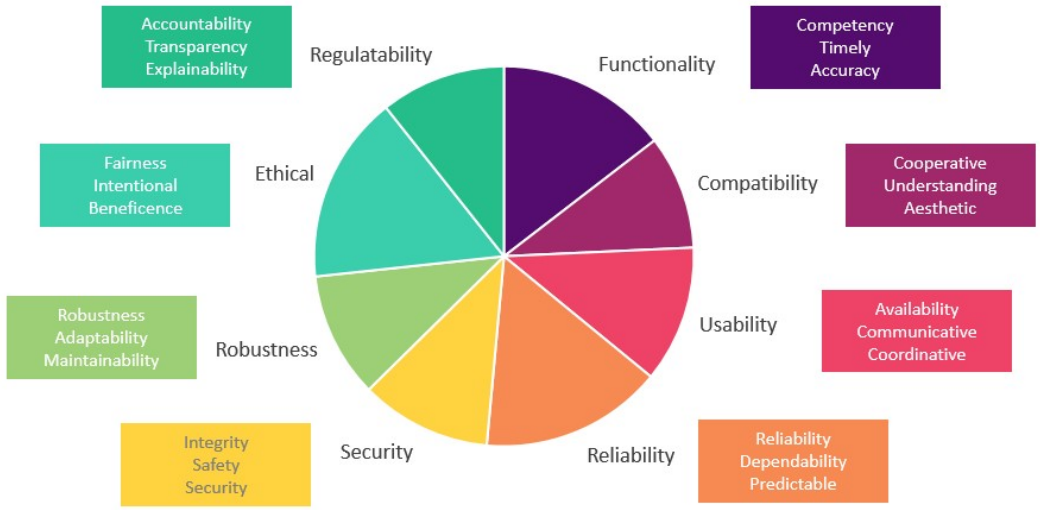


Fig. 1: Analysis of trust quality terms in the literature placed into categories, breakout box shows most cited words from each category.

The *concept paper* by the *European Union Aviation Safety Agency (EASA)* provides its first usable guidance for level 1 (human assistance) safety-related machine learning applications [16]. This guidance provides a roadmap to create a framework for AI trustworthiness ([16], pg. 8). The framework describes three techniques for analysing trustworthiness (safety, security and ethics-based), which are linked with the ethical guidelines developed by the EU commission (accountability, robustness, safety, oversight, privacy and data governance, non-discrimination and fairness, transparency, and societal and environmental well being).

The *DEpendable and Explainable Learning (DEEL) white paper* aims to identify challenges in the certification of systems using machine learning and to define a set of high-level properties for that purpose, such as auditability, data quality, explainability, maintainability, resilience, robustness, specifiability and verifiability ([44], pg. 22–23).

The *Aerospace Vehicle System Institute (AVSI) report* on machine learning summarises their findings on safety and certification aspects of emerging machine learning technologies that are applied to safety-critical aerospace systems [2]. This report provides several recommendations with respect to robustness, safety assurance, runtime assurance and interpretability when using machine learning in safety-critical applications.

The *UL 4600 standard* guides a user through the development of safety cases for fully automated vehicles (i.e. vehicles with no driver or supervisor) [50]. It is more of a standard of care and not a procedure for certification of fully automated vehicles. The *Laboratoire National de Métrologie et d'Essais (LNE) certification* [40] is a quality assurance standard for machine learning processes. The aim is to provide guidance for an applicant when obtaining certifications for their design, development, evaluation and maintenance in operational conditions.

#### 2.2.1 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 [28]. 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].

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 [59].

## 3 Key Considerations for Trustworthy AS

...

### 3.1 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 [14] which can be assessed using an appropriate risk assessment process from engineering, e.g. FMEA or sociotechnical disciplines [43]. 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.

#### 3.1.1 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.

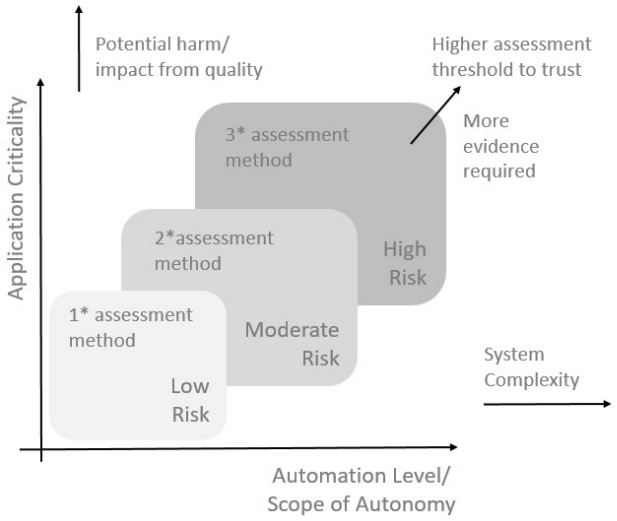


Fig. 2: Application criticality and automation level.

### 3.2 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 [54]. 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 [18]. 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 co-existence and harmony, are aspects of automation that would naturally extend the scope of the system.

#### 3.2.1 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

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. |

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 4).

### 3.3 Trustworthy Metrics

It is not sufficient for systems to just be trustworthy, they must be clearly identifiable as such. By developing useful and informative metrics specific to each trustworthy property, developers, verification engineers, regulators and end users can be better informed of the systems trustworthiness.

At the system level, 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 [52]. 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 [61]. 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 [21]. Kaur et al. suggest explainability metrics based on the euclidean distance between the system output compared to a panel of experts [32]. trustworthiness of computer systems using metrics designed to assess security, trust, resilience and agility [12]

In addition to system level trust, there must be metrics that provide transparency on specific trustworthy qualities. 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 [7].

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, 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. Other, more subjective qualities will use metrics appropriate for the discipline. Qualities such as fairness, cooperation and beneficence may require questionnaires, polls and user feedback to get the correct level of detail. Pete and Arianna - can you add to this last point?

### 3.4 Conflicting Qualities

A trustworthy system could be defined as having sufficient evidence that provides assurance against all trustworthy qualities that are deemed necessary and appropriate for the intended application. But what if maximising one system quality meant reducing another? For example, a self-driving vehicle has a performance quality to arrive efficiently at a destination by minimising journey time and is hence optimal at some non-zero speed.

ref assertions paper

## 4 Assessment Framework Vision

Below are the stages proposed for AS trustworthiness assessment process. From reviewing the literature, a com-

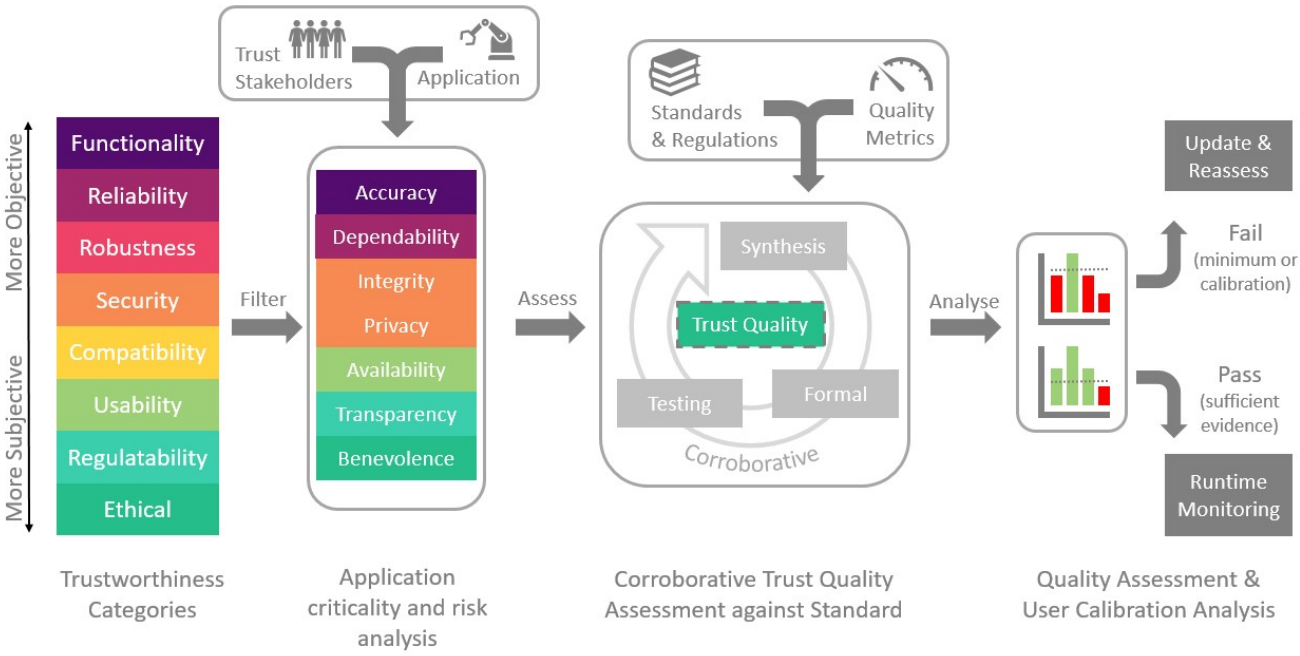


Fig. 3: AS trustworthiness assessment process

prehensive ontology has been derived based on trustworthy AS qualities. This index of qualities have been categorised and this can be used as a reference to begin the assessment.

### 4.1 Stage 1: Application Filtering & Criticality Analysis

Not all TAS qualities are relevant to all applications, so the first stage of the process is to identify the qualities that are relevant to the specified application and filter out those that are irrelevant or inappropriate. A criticality assessment is applied at this stage, using a process akin to a hazard analysis for each of the selected qualities not forgetting to assess the impact that automation scope may have on the final outcome, and what potential harm or impact of failing to uphold the specific quality will have on the outcomes of the system and it's users.

The consensus on which qualities to include and how important they are, will be decided by the *trust stakeholders*, a collection of individuals, groups and organisations who may include regulators, developers, end users and incidental users for example. Important factors that the stakeholders will consider are the scope of the automation; how automated the system is or the ambitions it has to be in the future, and the application criticality including the potential harm from the system not upholding the trustworthy attribute or quality. This should result in a list of qualities for the given application each with an assigned criticality rating, e.g. high, medium and low risk, that will be used in the assessment stage.

### 4.2 Stage 2: Identify Standards and Metrics

Stage 2 of the process is where each quality is assessed according to the appropriate standards and regulations relevant to that property, or the specifications on expected behaviour if such standards are immature or missing. At this point quality metrics should be considered that clearly visualise each of the trustworthiness qualities under scrutiny. Many metrics already exist, e.g. precision and recall for predictive analytics, but some will need to be generated and agreed upon by the community, e.g. cooperation, fairness. This is where the verifiability of a trustworthy quality, when thought about at the design stage, can pay off in the later assessment and verification stages, such as designing in status indicating LEDs into consumer electronics to visualise operational trustworthiness.

### 4.3 Stage 3: TASQ Assessment with Corroborative V&V

Stage 3 assesses each of the qualities from stage 1 using one or more verification approaches. Kress-Gazit et al. state that assessing the trustworthiness 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 *testing* methods [38]. Initially each quality will be assessed using either synthesis, testing or formal methods (runtime monitoring is considered in stage 4). However, the higher the risk criticality then a greater number of methods should be used. For example, a quality deemed low risk may use a single method to provide sufficient evidence of trustworthiness, but for higher risk categories corroborative evidence from multiple approaches is recommended. During each assessment, the quality will be verified against the respective set

of standards, regulations or specifications as appropriate and the quality metrics will be used to visualise the outcomes. For example, a robot swarm may be monitored for efficiency by monitoring idle individuals and representing this to the tester, regulator or end user as a utilisation score.

### 4.4 Stage 4: Assessment Analysis

Corroborative, mutually consistent evidence from diverse methodologies provides assurance that is greater in quality than evidence from single sources. The assessment analysis should inspect the evidence for each trust quality relative to the risk level based on the application criticality and automation scope. Confidence in the trustworthiness of the system will require more evidence for applications where there is a greater risk level. Where appropriate, there may be minimum requirement levels that must be achieved to attain trustworthiness for certain quality categories, for example 98% accuracy for an image recognition task.

Evidence that fails to corroborate, although does not fail in itself, e.g. sampling different parts of the input space, may be considered equivalent to single-source evidence and therefore, although not a failure, can only provide evidence for lower risk applications.

### 4.5 Stage 5: Visualising Trust

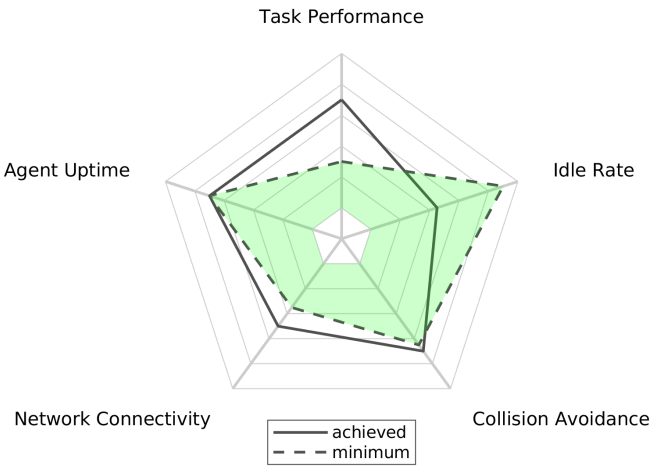


Fig. 4: Example of visualising trust in a fictitious swarm robot application.

Visualising the trustworthiness of the system will be an important aspect of not just presenting objective evidence to developers and regulators, but also to end users who may have a lay understanding of the technical aspects of the system. Discussion on what type of information to present to users and how this is done extends beyond the scope of this paper, but must be considered as this forms the reciprocal part of the trust agreement between user and system. However, there are several aspects of visualising trust that should be considered; which categories to display and the format, how the system performed during assessment or currently (if runtime monitoring is used), and a way of indicating the minimum requirements for each quality.

A fictitious example of how trustworthiness of a swarm application could be presented is shown in Fig.4, showing a radar plot of disparate qualities plotted on shared axes. It is clear at a glance that most of the system qualities deemed important, by the trust stakeholders for example, are at or above the minimum threshold required from the associated standards. Each quality can be seen compared to a minimum level (dashed line) and where the quality exceeds (Task Performance) or fails to reach a minimum level (Idle Rate) alerting the user to areas that may require remediating action (in this case the swarm is under-utilised and scores poorly for idle rate requiring a reduction in swarm size).

#### 4.5.1 Visualising Overall Trust

A complex, detailed visualisation of trust, as in Fig.4, may be suitable in some cases but at other times it may be more appropriate to present an abstract, higher level visualisation, especially for general public and lay users who are not operators, e.g. purchasing manager. 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*. Much like an internet browser may convince users of their privacy on a web page through some simple means, e.g. padlock icon, a similar approach could be taken with AS applications.

A vision of this rating system is presented in Table 2. A star rating is awarded based on the level of corroborative evidence from independent assessment methods that meet the appropriate compliance thresholds when viewed cumulatively across the spectrum of required qualities for the system. Essentially the more diverse the evidence the greater the trustworthiness rating will be, assuming that assessment outcomes are positive.

The proposal is that applications with lower risk levels are not required to provide the same level of evidence as higher risk applications, and therefore the highest burdens of proof are required exclusively for higher risk applications.

### 4.6 Stage 6: Assessment Outcomes

The assessment and subsequent analysis should identify qualities that are sufficient and deemed trustworthy and those that fail to meet the requirements.

#### 4.6.1 Insufficient Evidence: Update & Reassessment

If the analysis shows that the quality fails to meet the minimum specification or there is insufficient evidence that the quality meets the specification, the systems fails on that quality. The system must then pass into an iterative development cycle, where the system is updated and then reassessed.

#### 4.6.2 Sufficient Evidence: Monitoring

If there is sufficient evidence that the quality meets the specifications with respect to it automation scope and criticality, then the system passes that quality. In some applications it may be prudent to enter into a phase of operational assessment where metrics are observed in real-time. This runtime monitoring allows system trustworthiness to continue to be assessed which may be required if the system is adaptive to new environments or has evolving functionality.

### 4.7 Future Challenges in Standards

Riaz et al. [51] 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. [25], 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 [60] but must be translated to a computer readable format to act as an appropriate standard, or set of assertions [23], 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.

McDermid paper on ethical assurance ...

### 4.8 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 [36].

## 5 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] AFE 87 Project Members. *AFE 87: Machine learning.* Final Report 87-REP-01. Aerospace Vehicle Systems Institute, June 2020.

[3] R. Ashmore, R. Calinescu, and C. Paterson. "Assuring the machine learning lifecycle: Desiderata, methods, and challenges". In: *ACM Comput. Surv.* 54.5 (May 2021). URL: https://doi.org/10.1145/3453444.

[4] 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. 2012, pp. 303–307.

[5] 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.

[6] A. Avizienis et al. "Basic concepts and taxonomy of dependable and secure computing". In: *IEEE transactions on dependable and secure computing* 1.1 (2004), pp. 11–33.

[7] 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.

[8] J.-F. Bonnefon, A. Shariff, and I. Rahwan. "The social dilemma of autonomous vehicles". In: *Science* 352.6293 (2016), pp. 1573–1576. URL: https://www.science.org/doi/abs/10.1126/science.aaf2654.

[9] British Standards Institute. *BS8611:2016 Robots and robotic devices, guide to the ethical design and application of robots and robotic systems.* Online. 2016. URL: https://knowledge.bsigroup.com/products/robots-and-robotic-devices-guide-to-the-ethical-design-and-application-of-robots-and-robotic-systems/standard.

[10] 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.

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

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

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

[14] DIN. "Din spec 92001-1 Artificial Intelligence - Life Cycle Processes and Quality Requirements". In: (April 2019), pp. 1–23.

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

[16] European Aviation Safety Agency. *EASA concept paper first usable guidance for level 1 machine learning applications – Proposed issue 01.* Online. Apr. 2021. URL: https://www.easa.europa.eu/downloads/126648/en.

[17] M. Fewster and D. Graham. *Software Test Automation Effective use of test execution tools.* 1999, p. 570. URL: http://www.acm.org.

[18] M. Fisher et al. "Towards a framework for certification of reliable autonomous systems". In: *Autonomous Agents and Multi-Agent Systems* 35.8 (2021), p. 65.

Table 2: Trustworthy Autonomous System 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 |


[19] 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.

[20] 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.

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

[22] 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.

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

[24] R. Hawkins et al. *Guidance on the assurance of machine learning in autonomous systems (AMLAS)*. Guidance Version 1.1. University of York, Mar. 2021.

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

[26] IEEE Standards Association. *7007-2021 – IEEE ontological standard for ethically driven robotics and automation systems*. Online. 2021. URL: https://standards.ieee.org/ieee/7007/7070/.

[27] IEEE Standards Association. *7010-2020 – IEEE recommended practice for assessing the impact of autonomous and intelligent systems on human well-being*. Online. 2020. URL: https://standards.ieee.org/standard/7010-2020.html.

[28] International Organization for Standardization. *ISO/IEC/IEEE 29119 Software and systems engineering — Software testing*. Online. 2013. URL: https://www.iso.org/standard/45142.html.

[29] Y. Jia et al. *The role of explainability in assuring safety of machine learning in healthcare*. 2021.

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

[31] F. Kaakai et al. "Toward a machine learning development lifecycle for product certification and approval in aviation". In: *SAE Int. J. Aerosp.* 15.2 (May 2022).

[32] 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.

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

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

[35] 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.

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

[37] 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.

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

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

[40] LaboratoireNational de Metrologie et d'Essais. *Certification standard of processes for AI. Design, development, evaluation and maintenance in operational conditions*. Certification Standard 2. LaboratoireNational de Metrologie et d'Essais, July 2021.

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

[42] 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.

[43] C. Macrae. "Learning from the failure of autonomous and intelligent systems: Accidents, safety, and sociotechnical sources of risk". In: *Risk analysis* (2021).

[44] F. Mamalet et al. *White paper machine learning in certified systems*. Research Report hal-03176080. IRT Saint Exupery ANITI, 2021.

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

[46] 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.

[47] P. K. Murukannaiah et al. "New foundations of ethical multiagent systems". In: *Proc. of the 19th International Conference on Autonomous Agents and MultiAgent Systems*. AAMAS '20. Auckland, New Zealand: International Foundation for Autonomous Agents and Multiagent Systems, 2020, pp. 1706–1710.

[48] H. Nissenbaum and H. Daniel. "TrackMeNot: Resisting surveillance in web search". In: (2009).

[49] Z. Porter, I. Habli, and J. McDermid. "A Principle-based Ethical Assurance Argument for AI and Autonomous Systems". In: (2022), pp. 1–39. arXiv: 2203.15370. URL: https://arxiv.org/abs/2203.15370v1.

[50] D. Prince and P. Koopman. *UL 4600 technical overview*. Online. Oct. 2019. URL: https://ulse.org/UL4600.

[51] 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.

[52] 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.

[53] J. Rushby. "Runtime certification". In: *Runtime verification*. Ed. by M. Leucker. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 21–35.

[54] SAE International. *SAE J3016_201806 – taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles*. Online. 2018. URL: https://www.sae.org/standards/content/j3016_202104.

[55] M. Schwammberger et al. "Integrating Formal Verification and Simulation-based Assertion Checking in a Corroborative V&V Process". In: *arXiv preprint arXiv:2208.05273* (2022).

[56] 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.

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

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

[59] Trustworthy System Lab. *TAS-Verif*. https://github.com/TSL-UOB/TAS-Verif. Accessed: 2022-08-22.

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

[61] 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.

[62] 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.

[63] A. Winfield et al. "IEEE P7001: A proposed standard on transparency". In: *Frontiers in robotics and AI* 8 (2021), p. 225.

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

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

[66] 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.