A Conceptual Model of Operationalizable Trustworthiness for Autonomous Systems

Bálint Máté*, Diego Perez-Palacin†‡, Vincenzo Scotti†, Raffaela Mirandola†
*FZI Research Center for Information Technology, Karlsruhe, Germany
mate@fzi.de
†KASTEL, Karlsruhe Institute of Technology, Karlsruhe, Germany
{vincenzo.scotti, raffaela.mirandola}@kit.edu
‡Linnaeus University, Växjö, Sweden
diego.perez@lnu.se

***Abstract*—Autonomous systems are increasingly deployed across critical domains, yet research on trustworthy autonomous systems (TAS) appears to lack an agreed-upon, operationalizable taxonomy for trustworthiness characteristics. While trustworthy artificial intelligence (TAI) frameworks offer measurable characteristics with more established metrics, TAS literature tends to be fragmented with high-level models that can be difficult to apply in practice. This paper investigates the extent of alignment in TAS trustworthiness research and identifies potential sources of misalignment, including differences in system definitions, autonomy properties, stakeholder perspectives, and domain requirements. Based on this analysis, we synthesize a foundational two-pillar core of trustworthiness, the system's ability to dependably do what it is intended to do and the extent to which stakeholders can assess and justify confidence in this behavior. We propose an initial operationalization of this core using characteristics adapted from established TAI frameworks for which metrics and evaluation methods exist, and outline how it can be applied to specific application contexts. An example shows how stakeholder analysis guides characteristic adaptation.**

***Index Terms*—Trustworthiness, Autonomous Systems, Trust.**

I. Introduction

The rapid increase in computing power has enabled the development and deployment of autonomous systems (ASs) across various domains. Autonomous vehicles navigate complex traffic scenarios, robots perform intricate manipulation tasks, smart grids balance distributed energy resources, and agents coordinate to optimize complex processes. All of these systems function independently and perform tasks without explicit human control, for which they often utilize capabilities associated with artificial intelligence (AI) such as problem-solving, decision-making, and learning [1], [2].

In contrast to automation, where the system behavior is generally fully deterministic, autonomous systems often rely on perception, learning, and self-adaptation, which makes their behavior less predictable. Since the system lacks continuous human supervision, users must judge to what extent they can depend on the system and place an appropriate amount of trust in it. Otherwise they may avoid it entirely or rely on it excessively [3], [4]. The complex behavior of autonomous systems makes it difficult to assess the system behavior in certain scenarios [1]. Moreover, as systems learn and adapt, additional run-time validation and verification may be required [5].

Many autonomous capabilities are realized through AI and machine learning components, which trade interpretability for performance [2]. While these components enable autonomous behavior, they introduce non-determinism and opacity, hindering system understandability.

Research on trustworthy autonomous systems (TAS) is diverse, with numerous taxonomies originating from social sciences [3], [4], [6]. However, TAS literature appears to lack a generally accepted taxonomy with operationalizable trustworthiness characteristics. In contrast, work on trustworthy AI (TAI) has progressed more rapidly, with widely adopted risk management and trust frameworks and taxonomies that offer measurable characteristics [7], [8], [9], [10].

Significant work has been done to identify relevant trustworthiness characteristics in classical information and communications technology (ICT) systems [11], AI systems [7], [12], [13], and cyber-physical systems (CPSs) [14], [15], [16]. Contributions targeting autonomous systems exist [2], [4], [17], [18], [19], [20], but tend toward higher-level models that lack established metrics and methods for influencing trustworthiness characteristics. Combining them often yields abstract views with conflicting trust models. To our knowledge, no widely accepted taxonomy for TAS trustworthiness characteristics with associated metrics exists.

This work addresses the following research questions:

RQ1: To what extent does a unified view in academia exist on what a trustworthiness taxonomy for TAS must encompass?

RQ2: What are the sources of misalignment in the scientific community on TAS trustworthiness?

RQ3: How could a unified generally applicable TAS trustworthiness taxonomy be agreed upon, for which evaluation methods and metrics exist?

This work contributes: (i) an analysis of TAS trustworthiness views revealing the extent of alignment and divergence in the scientific community; (ii) an identification of four sources of divergence in TAS trustworthiness research; (iii) a foundational and extensible conceptual model of trustworthiness for autonomous systems, synthesized from prominent literature, with an initial operationalization to support the community in incrementally developing a unified TAS trustworthiness

taxonomy; (iv) an illustrative example applying the proposed conceptual model to a robot vacuum cleaner.

This paper is structured as follows: Section II establishes background. Sections III to V address **RQ1**, **RQ2** and **RQ3** respectively. Section VI concludes.

II. Background

Autonomous systems are designed to function without explicit human control. The complexity that comes with the independent behavior makes them less transparent to humans [1]. Furthermore, many autonomous capabilities are realized through AI components that introduce non-determinism and opacity [2].

Research on trustworthiness spans many domains. ICT provides foundational concepts (reliability, safety, security) [11]. Cyber-physical systems add physical-world interaction with timing constraints and real-time requirements [15]. Autonomous systems are characterized by independent operation, while AI systems by their computational method [7]. These domains overlap but are distinct: a system may be autonomous without AI (e.g., rule-based autopilot) or use AI without being autonomous (e.g., recommendation engine).

Analogously, trust and trustworthiness are related but distinct concepts. *Trust* is a psychological state of the trustor, described as willingness to depend on a system [3] or confidence that it will work as intended [21]. *Trustworthiness*, in contrast, is a property of the trustee (the system) that engenders trust [4], [22]. Thus, a system may be technically trustworthy yet fail to inspire trust if users cannot assess its capabilities. Conversely, users may trust a system that is not actually trustworthy. The relationship between trust and trustworthiness is dependent on the stakeholders' ability to assess the system's capabilities [23]. This consideration is especially relevant for domains where complexity and opacity hinder the assessment, such as autonomous systems.

Trust in technical systems exists within a sociotechnical context where both technical properties (reliability, security) and human factors (understanding, confidence) must be considered [12], [24]. Trust models from social sciences and human-automation interaction have been adapted for autonomous and AI systems [3], [6], [25]. We review these in Section III.

III. Existing Views on TAS Trustworthiness

This section examines the extent to which academia holds a unified view on what a trustworthiness taxonomy for TAS must encompass. Given our well-defined and unambiguous research question, we conducted a Rapid Literature Review (RLR) [26]. An RLR serves as an established, resource-efficient alternative to a Systematic Literature Review (SLR) and has proven effective for extracting essential literature [27]. As pragmatic evidence synthesis methods, RLRs relax some of the procedural rigor of SLRs to shorten the review process. In software engineering, RLRs have nonetheless been found to support decision-making with valuable evidence. A manual database search with topic-related search terms and citation tracking yielded 112 candidate works. A large language model-assisted screening of 343 *IEEE Xplore* and *ACM DL* records matching "trust* AND autonom*" added 16 further AS-focused works, for 128 in total. Search strings, screening criteria, and the full work list are in the replication package[1].

A. High-Level Trust Models

Foundational work on trust originates from philosophy and organizational psychology. Trust has been characterized as a relation between a trustor and a trustee within a restricted scope [20], [28]. Mayer et al. identified Ability, Benevolence, and Integrity as key pillars of trustworthiness [6]; Lee and Moray proposed the Purpose-Process-Performance framework for trust in automation [3], [29]; and Hoff and Bashir distinguished dispositional, situational, and learned trust [25]. These models show that trust in technical systems depends on factors beyond system properties alone, including the perceived benevolence of designers, trustor characteristics such as general propensity to trust and domain expertise, and contextual factors such as perceived risk [3], [20], [25].

The literature thus divides into works examining the psychology of trust in human trustors [3], [6], [12], [25] and works identifying technical system properties believed to influence trust [7], [11], [15]. This reflects the critical distinction between *trust* and *trustworthiness*. Trustworthiness as a system property is meaningful only to the extent that it generates appropriate trust, and trust cannot be objectively "built into" a system [23]. Trust calibration, which aligns the trustor's trust with the system's actual capabilities [3], depends on stakeholders being able to assess trustworthiness characteristics, motivating the operationalizable taxonomies surveyed next.

B. Survey of Trustworthiness Taxonomies

Having established that trust calibration requires assessable trustworthiness characteristics, we now review existing taxonomies that attempt to define such characteristics. We organize our review by domain, beginning with AS-specific work and proceeding to adjacent fields.

Several works directly address AS trustworthiness with varying scope. Devitt proposes a two-component model of competency and ethical integrity [4]; while reliability is argued as a core property, the subdivision into skills, motives, honesty, and character remains hard to quantify in engineering. Chance et al. identified eight trustworthiness categories from 60 papers and standards [19], though mixing in human-robot interaction and cyber-physical systems literature blurs autonomy-specific concerns. Cybersecurity, safety, health, and human-machine interaction have likewise been named as key trustworthiness factors for robots and autonomous systems [30].

Some taxonomies explicitly target autonomous cyber-physical systems. Four core concepts (trustworthiness, security, resilience, and dependability) have been proposed as driving a TAS taxonomy by Flammini et al. [2], though rather than offering a holistic evaluation approach, this work points

[1] doi.org/10.5281/zenodo.20341335

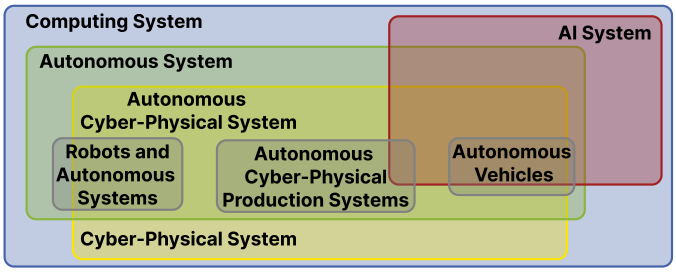

Fig. 1. Diversity of system scopes addressed in reviewed literature. Existing frameworks typically target specific subdomains (e.g., autonomous vehicles, robots) rather than autonomous systems in general.

towards relevant taxonomies from dependable computing, cybersecurity, and trustworthy AI domains.

Several works emphasize specific enablers of trustworthiness. Hoppe et al. argue that the integration of Ability, Benevolence, and Integrity is necessary but not sufficient, since systems must also demonstrate their trustworthiness, for which explainability is essential, and different stakeholders (passengers versus auditors) yield different requirements [20]. Verifiability has been positioned as key to trust, used both to assess reliability and to establish beneficiality [18]. Transparency has likewise been emphasized. Explanations do not improve dependability per se, but improve predictability and thus user trust [17].

Beyond AS-specific work, adjacent domains offer established taxonomies. CPS trustworthiness is well-researched, with six characteristics proposed by [14], surveys of up to 13 main attributes [16], and the NIST CPS Framework defining trustworthiness through security, privacy, safety, reliability, and resilience [15]. Trustworthy AI is an active area, with scoping reviews [9], [10], cognitive and emotional trust dimensions [12], and frameworks including the NIST AI RMF [7] (7 characteristics), the AI Trust Maturity Model [8] (7 characteristics), and the EU AI Act. Foundational ICT work centers on the dependability taxonomy [2], [11], [21], though the equivalence between trustworthiness and dependability weakens for learning systems, while ISO/IEC TS 5723 defines trustworthiness as "the ability to meet stakeholders' expectations in a verifiable way" [31]. These adjacent-domain taxonomies provide valuable foundations but do not fully address autonomous decision-making.

Fig. 1 illustrates the diversity of system scopes, showing that existing frameworks often target specific subdomains rather than autonomous systems in general.

C. Key Observations

Despite terminological differences, several themes recur across the different domains. Dependability is seen as a common requirement, and nearly all frameworks treat its sub-attributes reliability and safety as fundamental [2], [11], [15]. For autonomous systems specifically, the importance of transparency and explainability is widely acknowledged [32], [33], as users and regulators need to understand system behavior to calibrate trust appropriately. Security concerns, often in the form of protection against adversarial threats and confidentiality, also appear in each investigated domain [2], [7], [15], though with varying emphasis and exact threats. Finally, multiple authors recognize that learning and self-adaptive systems present verification challenges, requiring ongoing verification throughout the system's lifecycle rather than one-time certification as for classical systems [2], [5].

At the same time, the works differ substantially in how trustworthiness is conceptualized. They vary in granularity (from 5 characteristics [11] to over 40 [16]), in abstraction level [4], [7], in the boundary drawn between an "autonomous system", an "AI system", and a "cyber-physical system" [19], [30], and in stakeholder perspective [20]. Section IV examines these sources of misalignment in detail.

Finding on RQ1: TAS lacks a mature, operationalizable trustworthiness framework. While there is agreement on core themes such as dependability, transparency, and security, TAS-specific work tends toward abstract models. Operational taxonomies exist in adjacent domains (TAI, CPS, ICT, HAI) but do not fully address AS-specific concerns.

IV. Contextual Factors in TAS Research

Building on the divergences identified in the previous section, we now examine the contextual factors that shape how different works conceptualize TAS trustworthiness, and answer the question why existing approaches differ based on their intended scope, audience, and application context.

A. Target System Scope

A fundamental source of divergence is what each work considers an "autonomous system." The term is used for systems ranging from simple automated controllers to complex learning agents, and reviews scope it differently. Some encompass robots and cyber-physical systems alongside autonomous systems [30], others frame TAS as "autonomous CPS operating in critical environments" [2], and others span autonomous systems, human-robot interaction, and ML-enabled systems together [19]. As established in Section II, these domains are conceptually distinct. TAI work emphasizes fairness and bias mitigation, TAS research emphasizes concerns arising from reduced human oversight, and CPS literature emphasizes physical safety and real-time constraints [14]. When such domain-specific requirements are mixed without acknowledgment, the resulting taxonomies become difficult to apply consistently.

B. Emphasized Autonomy Properties

Even focusing specifically on autonomous systems, works emphasize different autonomy properties as most relevant for trustworthiness: (i) *delegated responsibility*: higher autonomy delegates more responsibility to the system, with accountability implications when things go wrong [34], [35]; (ii) *reduced human monitoring*: operating with little or no oversight raises the capability bar a trustor expects before depending on the system [35], [36]; (iii) *opacity*: from the complex behavior enabling autonomy [1], [24] or from AI/ML components [22],

[37], making the system harder to understand and predict; and (iv) *behavioral adaptation*: learning and self-adaptive systems change over time, requiring dynamic verification rather than one-time certification [5], [24].

C. Varying Stakeholder Perspectives

Trustworthiness has different meaning to different stakeholders [33]. User-focused works examine how humans perceive and form trust [4], engineering-focused frameworks emphasize measurable technical parameters [7], [20], and regulatory perspectives emphasize legal alignment such as the EU AI Act and ISO/IEC standards [9]. These perspectives yield taxonomies with different characteristic priorities and abstraction levels, complicating direct comparison. Trust is also transitive: a user's trust in a system depends partly on trust in the institutions behind it, which taxonomies rarely address explicitly.

D. Criticality and Domain Specificity

The required rigor of trustworthiness assurance varies with application criticality. Safety-critical automotive, aeronautical, or medical systems require verification unnecessary for less critical applications [17], yielding domain-specific taxonomies (Fig. 1). The applicability of foundational taxonomies, such as dependability [11] or AI trustworthiness [7], also depends on system architecture and deployment context [2], since a system with AI components must address fairness, bias, and accuracy, while a rule-based one need not. Such architectural differences are often unaccounted for in general-purpose taxonomies.

Finding on RQ2: The key contextual factors shaping the conceptualization of TAS trustworthiness include: (1) the exact scope of what constitutes an autonomous system, (2) the emphasized autonomy properties, (3) different stakeholder perspectives, and (4) criticality and domain-specificity.

V. Towards a Generally Applicable TAS Taxonomy

Having identified sources of misalignment, we now propose a conceptual model for constructing generally applicable TAS trustworthiness taxonomies for which metrics and evaluation methods exist.

Given the lack of operationalizable TAS-specific taxonomies identified in Sections III and IV, we propose a framework that combines a TAS-specific core with operationalizable characteristics from mature TAI frameworks.

A. Core TAS View: The Two Pillars

Despite the divergences identified earlier, we observe that TAS and ICT literature converges on a common core view. To represent this, we propose a structure of **ability** and **assessability** pillars. While the term ability also appears in trust literature (e.g., the ABI model [6]), in our proposed model both pillars synthesize related concepts found across a multitude of sources. Table I shows how the trustworthiness framings of works across the reviewed literature map onto these two pillars. Each work is coded in its own terms, and the mapping is an interpretive coding offered to make the synthesis inspectable rather than as proof of convergence.

This two-pillar structure captures the essence of TAS trustworthiness. A system must both have the technical capability to perform correctly and reliably (ability) and *demonstrate* its trustworthiness by endowing stakeholders with the possibility to evaluate, verify, or have justified confidence in said system behavior (assessability).

Some perspectives extend this view with benevolence, integrity, or beneficiality dimensions. However, these concepts originate from human trust models and do not translate directly to technical systems. A system cannot "intend" good or "be honest" in the way humans can. Where relevant, these concerns can be addressed through the two pillars (e.g., integrity as validity and transparency or benevolence as beneficial design intent, assessed through transparency).

B. Operationalization

We propose specifying each pillar using characteristics from mature TAI frameworks, which offer the granularity needed for practical application. Nevertheless, their interpretation requires adjustment for AS contexts. We select the NIST AI RMF [7] as the primary source for the set of characteristics for the two pillars due to the framework's wide adoption and comprehensive characteristic coverage. We decompose its compound characteristics, such as *secure and resilient*, into atomic ones, so that evaluation methods and metrics attach at this finer grain. Fig. 2 presents the resulting conceptual model. Table II presents methods and metrics for evaluating both pillars. Metrics for assessability characteristics are less standardized than for ability. Gaps remain in dynamic verification and autonomy-aware assessment.

The ability pillar is rooted in classical dependability [11] and comprises five characteristics. *Validity* captures whether the system performs as intended and fulfills its requirements [7]; for autonomous systems, this must hold continuously as the system adapts and encounters unexpected inputs, requiring dynamic verification [5]. *Reliability* refers to the continuity of correct service, where the independent operation of autonomous systems amplifies the consequences of failures. *Safety* denotes the absence of catastrophic consequences on users or environment, particularly critical for embodied autonomous systems such as vehicles and robots. *Security*, the composite of confidentiality, integrity, and availability, becomes harder to ensure as the attack surface grows with each sensor and component while human supervision is reduced. Finally, *resilience* [40] captures the persistence of dependability under change, important whenever autonomous systems must handle environmental shifts and component failures with limited oversight.

While the NIST AI RMF includes privacy and fairness as core characteristics, these differ in kind from the two pillars. The pillars are *execution-level*: they relate a system's behavior to its intended behavior (ability) and make that relationship

TABLE I

MAPPING OF SELECTED TRUSTWORTHINESS FRAMINGS ONTO THE *ability* AND *assessability* PILLARS, RECORDED IN EACH SOURCE'S OWN TERMS. A BLANK CELL MARKS A SINGLE-PILLAR SCOPE, NOT THE REJECTION OF THE OTHER. *Linked* RECORDS WHETHER THE FRAMING IS CAST AS KEY TO *trust* OR TO *trustworthiness*.

Work	Domain	Linked	Ability framing	Assessability framing
[3]	HAI	trust	performance (reliability, predictability)	process (understandable algorithms)
[38]	AS	both	reliability, robustness	transparency, V&V, certifiability
[34]	AS	trust	—	accountability
[33]	AS	both	—	transparency (IEEE P7001)
[23]	AS/AI	trustw.	dependably does what intended	assurance cases
[17]	AS	both	dependability	transparency, explanations
[39]	AS	trust	reliability	understanding of how decisions are derived
[19]	AS	trustw.	reliability, accuracy	verifiability, transparency (IEEE P7001)
[18]	AS	both	reliability	verifiability
[24]	AS	trustw.	behaves as expected despite faults	verification, explainability
[2]	AS	both	dependability, security, resilience	transparency, explainability
[21]	AS	trust	dependability, antifragility	assessed confidence in requirement fulfillment
[20]	AV	both	relevant capabilities	explainability
[7]	AI	trustw.	validity, reliability, safety, security, resilience	transparency, explainability, interpretability, accountability
[8]	AI	both	robustness, safety	explainability, transparency, accountability
[37]	AI	trust	reliability, robustness	interpretability
[15]	CPS	trustw.	security, safety, reliability, resilience	—
[16]	CPS	trustw.	assured correct performance	transparency, provability (evidence)
[11]	ICT	trust	dependability: justifiably trusted service	justification via fault removal and forecasting
[31]	ICT	trustw.	meet stakeholder expectations	verifiability, measurability, demonstrability

HAI = Human-Automation Interaction; AS = Autonomous Systems; AI = Artificial Intelligence; AV = Autonomous Vehicles; CPS = Cyber-Physical Systems; ICT = Information and Communications Technology.

evaluable (assessability), while staying agnostic about whether the intended behavior is itself desirable. Privacy, fairness, and ethical concerns are instead *specification-level*: they constrain what the intended behavior *should* be. They are therefore not additional pillars but context-specific requirements: once a context adopts one (e.g., privacy when autonomous systems collect personal data, fairness when their decisions differently affect people), conformance to it becomes a validity (ability) concern and its demonstration an assessability concern. Such characteristics are introduced during the customization step (Section V-C) when the application context demands it.

The assessability pillar covers four characteristics that enable stakeholders to evaluate the ability pillar. Their definitions are not unified across the community [13], [32]. *Transparency* is the extent to which information about the system and its outputs is accessible to stakeholders [7], [32]; for autonomous systems this includes decision-making processes, system boundaries, and limitations, with differing needs across users, engineers, and incident investigators [33]. *Explainability* refers to methods that rationalize and clarify system behavior [32], [41], critical for autonomous systems with opaque AI components both at operation time and post-hoc. *Interpretability* is the capacity to derive meaning from system behavior [32], [42], especially relevant in regulated domains. *Accountability* is the a-posteriori mechanism linking events to a natural or legal person [43]. For component-based autonomous systems, it requires clear attribution of causality across autonomous decision chains [44].

C. Application Example

The two-pillar model (Fig. 2) serves as a baseline for constructing application-context-specific trustworthiness taxonomies. Deriving such a taxonomy involves analyzing the system context, adapting the default characteristics, selecting evaluation methods, and defining fulfillment levels. A general customization workflow over these steps is left to future work. Here we illustrate them on a robot vacuum cleaner, a system whose application domain is autonomous indoor floor cleaning, navigating rooms, avoiding obstacles, returning to the charging dock, and including cameras for navigation and home monitoring in some models.

TABLE II

METHODS AND METRICS FOR TRUSTWORTHINESS CHARACTERISTICS.

Char.	Methods	Metrics
Ability		
Validity	Formal verification, conformance testing, validation datasets	Testing coverage, conformance scores, OOD detection rates
Reliability	Statistical testing, concept drift detection	MTTF, MTBF, failure rate, prediction confidence
Safety	Hazard analysis, fault tree analysis, safety cases	SIL (IEC 61508), ASIL (ISO 26262), DAL (DO-178C)
Security	Penetration testing, threat modeling, red teaming	Vulnerability counts, attack success rate, adversarial robustness
Resilience	Fault injection, stress testing, chaos engineering	Recovery time, degradation levels, fault coverage
Assessability		
Transp.	Documentation review, stakeholder interviews	IEEE P7001 transparency levels
Explain.	LIME, SHAP, attention visualization, user studies	Fidelity, consistency, comprehensibility scores
Interpr.	Task-based evaluation, user studies	User understanding scores, prediction accuracy
Account.	Causal tracing, decision logging, provenance tracking	Audit trail completeness, logging coverage

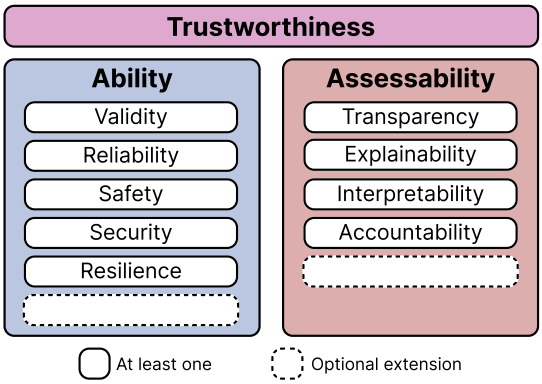

Fig. 2. Two-pillar TAS trustworthiness taxonomy with operationalizable characteristics.

1) Analyze System Context: We apply IEEE P7001's stakeholder classification and consider two groups whose priorities differ substantially. For *users*, trust centers around the system doing its job reliably (cleaning effectively, returning to dock, recovering when stuck) and, for camera-equipped models, not leaking home data; users largely do not consider safety (no hazardous components), security, or accountability, and on the assessability side want simple status indicators and explanations rather than detailed logging. *Certification agencies* instead require compliance evidence for validity and reliability (e.g., MTTF documentation), basic electrical safety per IEC 60335, security analysis, and GDPR compliance; for assessability, they expect P7001-compliant logging, cleaning history records, and accountability through data access audit trails. This comparison reveals that safety, security, and accountability may not be directly relevant for user trust but are essential for certification.

2) Adapt Characteristics: Table III shows the adaptation of characteristics for users (and the corresponding methods and metrics from the following step). Safety, security, and accountability are removed since they are not directly relevant for user trust in this context. Privacy is added as it is essential for camera-equipped models.

3) Select Evaluation Methods: The third column of Table III lists the corresponding methods and metrics for each adapted characteristic.

4) Define Fulfillment Levels: Discrete cumulative fulfillment levels are defined per characteristic (e.g., illustrative levels for validity coverage: L1 >80%, L2 >90%, L3 >95%, L4 >98%), enabling dimension-wise comparison across products. The full level definitions for all characteristics are provided in the replication package.

Stakeholders should be actively involved in customization, both to identify relevant per-characteristic interpretations and priorities and to validate that the resulting trustworthiness scores correlate with their perceived trust. While the conceptual model still requires empirical validation, we ground its operationalization in the widely adopted NIST AI RMF and illustrate its application on an example.

TABLE III
CHARACTERISTIC ADAPTATION, METHODS, AND METRICS FOR THE ROBOT VACUUM CLEANER (USER PERSPECTIVE).

Char.	Domain-Specific Adaptation	Methods and Metrics
Ability		
Validity	Complete floor coverage, correct room recognition, returns to the station	Coverage, room recognition accuracy, return home rate
Reliability	Consistent cleaning, does not get stuck, battery longevity	Mean Time Between Failures, cleaning consistency, rate of getting stuck
Resilience	Recovery from stuck situations, handles being stepped on or getting wet	Recovery success rate, recovery time
Privacy	Camera data must not leak, home map protection	GDPR compliance, data retention policies, map encryption
Assessability		
Transp.	Status indicators, push notifications	Status indicator clarity, notification reliability
Explain.	Reasons for returning to dock, getting stuck, missed areas	User comprehension scores for status messages
Interpr.	Visibility of room maps, cleaning schedules for users	Map accuracy, schedule predictability

Finding on RQ3: A foundational and extensible TAS trustworthiness conceptual model can be derived by combining the two-pillar core from TAS literature with operationalizable characteristics from mature TAI frameworks. This approach provides measurable properties while maintaining grounding in TAS-specific concerns. Stakeholder-specific instantiation and AS-specific considerations (dynamic verification, reduced monitoring, opacity) are necessary for practical application.

VI. CONCLUSION AND FUTURE WORK

This work investigated the state of trustworthiness research for autonomous systems. We found that a unified view on TAS trustworthiness taxonomies does not appear to exist. While literature agrees on core themes, major taxonomies are rather high-level, which makes it difficult to associate evaluation methods and appropriate quantification with their view on trust and trustworthiness. Building on existing literature, we started an effort to obtain an operationalizable trustworthiness taxonomy for autonomous systems. We harmonized existing work by proposing a two-pillar conceptual model with an initial operationalization, providing measurable characteristics for autonomous systems performing in an arbitrary application context, and illustrated its context-specific adaptation on an example.

Future work will develop a full customization workflow and validate the conceptual model through case studies in higher-criticality domains under the involvement of various stakeholder groups to assess the generalizability of our concept. Our work could also benefit from the establishment of a unified evaluation methodology for currently under-specified characteristics of the assessability pillar.

ACKNOWLEDGMENT

Some parts of our work were assisted by Generative AI. We used Grammarly for Education and Anthropic's Claude Opus 4.8 to help us refine the manuscript, and Qwen3 during the LLM-assisted literature review (also see Section III).

REFERENCES

[1] S. Shahrdar, L. Menezes, and M. Nojoumian, "A survey on trust in autonomous systems," in *Intelligent Computing*, vol. 857, Cham: Springer International Publishing, 2019, pp. 368–386.

[2] F. Flammini, C. Alcaraz, E. Bellini, S. Marrone, J. Lopez, and A. Bondavalli, "Towards trustworthy autonomous systems: Taxonomies and future perspectives," *IEEE Trans. Emerg. Topics Comput.*, vol. 12, no. 2, pp. 601–614, Apr. 2024.

[3] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Hum. Factors*, vol. 46, no. 1, pp. 50–80, 2004.

[4] S. K. Devitt, "Trustworthiness of autonomous systems," in *Foundations of Trusted Autonomy*, Cham: Springer International Publishing, 2018, pp. 161–184.

[5] A. Bombarda et al., "Evaluation framework for autonomous systems: The case of programmable electronic medical systems," *IEEE Trans. Softw. Eng.*, vol. 50, no. 4, pp. 995–1014, Apr. 2024.

[6] R. C. Mayer, J. H. Davis, and F. D. Schoorman, "An integrative model of organizational trust," *Acad. Manag. Rev.*, vol. 20, no. 3, pp. 709–734, 1995.

[7] E. Tabassi, "Artificial intelligence risk management framework (AI RMF 1.0)," National Institute of Standards and Technology (U.S.), Gaithersburg, MD, NIST AI 100-1, Jan. 2023.

[8] M. Mylrea and N. Robinson, "Artificial intelligence (AI) trust framework and maturity model: Applying an entropy lens to improve security, privacy, and ethical AI," *Entropy*, vol. 25, no. 10, p. 1429, Oct. 2023.

[9] A. Nastoska, B. Jancheska, M. Rizinski, and D. Trajanov, "Evaluating trustworthiness in AI: Risks, metrics, and applications across industries," *Electronics*, vol. 14, no. 13, 2025.

[10] S. Mehrotra, J. Huang, X. Fu, R. Dobbe, C. I. Sánchez, and M. de Rijke, "Understanding AI trustworthiness: A scoping review of AIES & FAccT articles," *J. Artif. Intell. Res.*, vol. 85, Mar. 2026.

[11] A. Avizienis, J.-C. Laprie, B. Randell, and C. Landwehr, "Basic concepts and taxonomy of dependable and secure computing," *IEEE Trans. Dependable Secure Comput.*, vol. 1, no. 1, pp. 11–33, Jan. 2004.

[12] E. Glikson and A. W. Woolley, "Human trust in artificial intelligence: Review of empirical research," *Acad. Manag. Ann.*, vol. 14, no. 2, pp. 627–660, Jul. 2020.

[13] A. Toney, K. Curlee, and E. Probasco, "Trust issues: Discrepancies in trustworthy AI keywords use in policy and research," in *Proc. ACM Conf. Fairness, Accountability, and Transparency (FAccT)*, Jun. 2024, pp. 2222–2233.

[14] H. A. Boyes, "Trustworthy cyber-physical systems — a review," in *Proc. 8th IET Int. Syst. Safety Conf*, Oct. 2013, pp. 1–8.

[15] E. R. Griffor, C. Greer, D. A. Wollman, and M. J. Burns, "Framework for cyber-physical systems: Volume 1, overview," National Institute of Standards and Technology, Gaithersburg, MD, NIST SP 1500-201, Jun. 2017.

[16] N. Gol Mohammadi, *Trustworthy Cyber-Physical Systems: A Systematic Framework towards Design and Evaluation of Trust and Trustworthiness*. Wiesbaden: Springer Fachmedien Wiesbaden, 2019.

[17] N. TaheriNejad, A. Herkersdorf, and A. Jantsch, "Autonomous systems, trust, and guarantees," *IEEE Des. Test*, vol. 39, no. 1, pp. 42–48, Feb. 2022.

[18] M. R. Mousavi et al., "Trustworthy autonomous systems through verifiability," *Computer*, vol. 56, no. 2, pp. 40–47, Feb. 2023.

[19] G. Chance, D. B. Abeywickrama, B. LeClair, O. Kerr, and K. Eder, *Assessing trustworthiness of autonomous systems*, 2023. arXiv: 2305.03411 [cs.AI].

[20] I. Hoppe, W. Hagemann, I. Stierand, A. Hahn, and A. Bolles, "Challenges for trustworthy autonomous vehicles: Let us learn from life," *Syst. Eng.*, vol. 27, no. 4, pp. 789–800, Jul. 2024.

[21] V. Grassi, R. Mirandola, and D. Perez-Palacin, "A conceptual and architectural characterization of antifragile systems," *J. Syst. Softw.*, vol. 213, p. 112051, Jul. 2024.

[22] P. Palazzolo, B. Stahl, and H. Webb, "Measurable trust: The key to unlocking user confidence in black-box AI," in *Proc. 2nd Int. Symp. Trustworthy Auton. Syst. (TAS)*, Sep. 2024, pp. 1–7.

[23] D. M. Tate, "Trust, trustworthiness, and assurance of AI and autonomy," Institute for Defense Analyses, Alexandria, VA, IDA Document D-22631, 2021.

[24] J. Sifakis and D. Harel, "Trustworthy autonomous system development," *ACM Trans. Embed. Comput. Syst.*, vol. 22, no. 3, pp. 1–24, May 2023.

[25] K. A. Hoff and M. Bashir, "Trust in automation: Integrating empirical evidence on factors that influence trust," *Hum. Factors*, vol. 57, no. 3, pp. 407–434, May 2015.

[26] B. Cartaxo, G. Pinto, and S. Soares, "The role of rapid reviews in supporting decision-making in software engineering practice," in *Proc. Int. Conf. Eval. Assess. Softw. Eng. (EASE)*, 2018, pp. 24–34.

[27] D. Amalfitano et al., "A research roadmap for augmenting software engineering processes and software products with generative AI," *ACM Trans. Softw. Eng. Methodol.*, Jan. 2026.

[28] A. Baier, "Trust and antitrust," *Ethics*, vol. 96, no. 2, pp. 231–260, Jan. 1986.

[29] J. Lee and N. Moray, "Trust, control strategies and allocation of function in human-machine systems," *Ergonomics*, vol. 35, no. 10, pp. 1243–1270, Oct. 1992.

[30] H. He, J. Gray, A. Cangelosi, Q. Meng, T. M. McGinnity, and J. Mehnen, "The challenges and opportunities of artificial intelligence for trustworthy robots and autonomous systems," in *Proc. 3rd Int. Conf. Intell. Robot. Control Eng. (IRCE)*, Aug. 2020, pp. 68–74.

[31] "Trustworthiness — vocabulary," International Organization for Standardization, Standard ISO/IEC TS 5723:2022, 2022.

[32] Z. C. Lipton, "The mythos of model interpretability," *Queue*, vol. 16, no. 3, pp. 31–57, Jun. 2018.

[33] A. F. T. Winfield et al., "IEEE P7001: A proposed standard on transparency," *Front. Robot. AI*, vol. 8, Jul. 2021.

[34] F. Galdon and S. J. Wang, "Addressing accountability in highly autonomous virtual assistants," in *Human Interaction and Emerging Technologies*, T. Ahram, R. Taiar, S. Colson, and A. Choplin, Eds., 2020, pp. 10–14.

[35] M. Attia, M. Hossny, S. Nahavandi, M. Dalvand, and H. Asadi, "Towards trusted autonomous surgical robots," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2018, pp. 4083–4088.

[36] F. Galdon, A. Hall, and S. J. Wang, "Designing trust in highly automated virtual assistants: A taxonomy of levels of autonomy," in *Artificial Intelligence in Industry 4.0*, vol. 928, Cham: Springer International Publishing, 2021, pp. 199–211.

[37] M. A. Langford, S. Zilberman, and B. Cheng, "Anunnaki: A modular framework for developing trusted artificial intelligence," *ACM Trans. Auton. Adapt. Syst.*, vol. 19, no. 3, 17:1–17:34, Sep. 2024.

[38] J. B. Lyons, M. A. Clark, A. R. Wagner, and M. J. Schuelke, "Certifiable trust in autonomous systems: Making the intractable tangible," *AI Mag.*, vol. 38, no. 3, pp. 37–49, Sep. 2017.

[39] Z. Assaad and C. Boshuijzen-van Burken, "Ethics and safety of human-machine teaming," in *Proc. 1st Int. Symp. Trustworthy Auton. Syst. (TAS)*, Jul. 2023, pp. 1–8.

[40] J. Laprie, "From dependability to resilience," in *Proc. IEEE Conf. Dependable Syst. Netw.*, 2008, G8–G9.

[41] P. Li, U. Siddique, and Y. Cao, "From explainability to interpretability: Interpretable policies in reinforcement learning via model explanation," Jan. 2025. arXiv: 2501.09858 [cs.LG].

[42] G. Vilone and L. Longo, "Notions of explainability and evaluation approaches for explainable artificial intelligence," *Inf. Fusion*, vol. 76, pp. 89–106, Dec. 2021.

[43] S. Kacianka, K. Beckers, F. Kelbert, and P. Kumari, "How accountability is implemented and understood in research tools: A systematic mapping study," in *Proc. Int. Conf. Product-Focused Softw. Process Improvement (PROFES)*, vol. 10611, Cham: Springer International Publishing, 2017, pp. 199–218.

[44] H. Schmidt, I. Poernomo, and R. Reussner, "Trust-by-contract: Modelling, analysing and predicting behaviour of software architectures," *J. Integr. Des. Process Sci.*, vol. 5, pp. 25–51, Aug. 2001.