F. Flammini, et al., "Towards Trustworthy Autonomous Systems: Taxonomies and Future Perspectives", *IEEE Transactions on Emerging Topics in Computing*, 2022.
http://doi.org/https://doi.org/10.1109/TETC.2022.3227113
NICS Lab. Publications: https://www.nics.uma.es/publications

# Towards Trustworthy Autonomous Systems: A Survey of Taxonomies and Future Perspectives

Francesco Flammini[1], Cristina Alcaraz[2], Emanuele Bellin[3], Stefano Marrone,[4] Javier Lopez[2], Andrea Bondavalli[5]

[1]School of Innovation, Design, and Engineering, Mälardalen University, Eskilstuna

[2]Computer Science Department, University of Malaga

[3]DHLab , University of Roma Tre, Rome

[4] Dep. of Mathematics and Physics, Univ. of Campania "Luigi Vanvitelli", Caserta

[5] Department of Mathematics and Informatics, University of Florence, Florence

francesco.flammini@mdu.se; alcaraz@uma.es; emanuele.bellini@ieee.org;
stefano.marrone@unicampania.it; javierlopez@uma.es;
andrea.bondavalli@unifi.it

**Abstract**

The class of Trustworthy Autonomous Systems (TAS) includes cyber-physical systems leveraging on self-x technologies that make them capable to learn, adapt to changes, and reason under uncertainties in possibly critical applications and evolving environments. In the last decade, there has been a growing interest in enabling artificial intelligence technologies, such as advanced machine learning, new threats, such as adversarial attacks, and certification challenges, due to the lack of sufficient explainability. However, in order to be trustworthy, those systems also need to be dependable, secure, and resilient according to well-established taxonomies, methodologies, and tools. Therefore, several aspects need to be addressed for TAS, ranging from proper taxonomic classification to the identification of research opportunities and challenges. Given such a context, in this paper address relevant taxonomies and research perspectives in the field of TAS. We start from basic definitions and move towards future perspectives, regulations, and emerging technologies supporting development and operation of TAS.

Keywords: Trustworthy Autonomous Systems, Dependability, Cyber-Resilience, Cybersecurity, Artificial Intelligence, Intelligent Systems

## 1 Introduction

The rapid technological evolution in Artificial Intelligence (AI) and the convergence of cyber and physical elements into Cyber-Physical Systems (CPS)

[1] have enabled a deeper integration of AI within complex systems that can take decisions and act without human intervention, possibly supported by other promising technologies such as Digital Twins (DT) [2]. This results in Autonomous Systems (AS), such as collaborative robots, which must be trustworthy when deployed in critical applications, such as transportation and production systems. For that reason, the area of Trustworthy Autonomous Systems (TAS) has emerged in recent years as a new paradigm and a key research domain involving multiple interdisciplinary communities.

TAS can be effectively presented from the perspective of CPS, which have evolved from the traditional class of embedded systems, where physical and cyber components are strictly interacting, analyzed and controlled by using holistic models and AI, possibly adopting. From such perspective, TAS are autonomous CPS operating in critical environments where failures can have serious consequences including loss of human lives. In this respect, TAS require a convergence of Infrastructure, Computing and Intelligence and the related key enabling technologies are represented by: :

- Infrastructure: the Internet of Things (IoT) or Internet of Everything (IoE) [3], which attracted many researchers working in the areas of Wireless Sensor Networks (WSN), connected and distributed systems;

- Computing: Edge-Fog-Cloud, the latter allowing engineers to leverage on computation resources well beyond the ones available in local devices, with a paradigm shift towards unprecedented scenarios of digital twinning, including predictive maintenance and proactive safety.

- Intelligence: Data Science and Machine Learning (ML), including Big Data analytics, which attracted and extended the interest of researchers in AI, as well as in many related areas, towards data-driven approaches;

The main inter-relations among the concepts of AI/ML, (autonomous) CPS, and (cognitive) DT are summarized in the class diagram of Fig. 1.

A set of new threats, vulnerabilities, and risks emerged with those new paradigms and technologies, due to systems growth into open and large System-of-Systems (SoS) [4], showing ubiquity, heterogeneity and pervasiveness of cyber-components, which became increasingly intelligent, adaptive, and evolving, also due to Software Over The Air (SOTA) dynamic updates, and hence more complex and less predictable. That has posed many challenges, especially for the certification of safety-critical systems featuring difficult to explain behaviors, such as those originated by Artificial Neural Networks (ANN) and Deep Learning (DL) [5]. One additional implication is on the evolution of standards and regulations to cope with intelligent, adaptive, and evolving systems in critical applications.

Although AI is mostly seen as a threat to safety, due to its possible unpredictability and new adversarial attacks it is vulnerable to, it is also true that AI enables self-protecting systems featuring on-line data-driven risk assessment for autonomous threat detection and counteraction; we will discuss some of

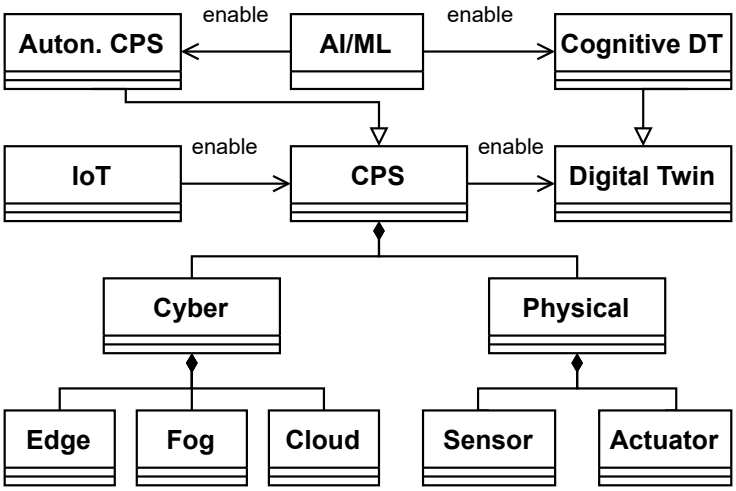


Figure 1: Class diagram for CPS, IoT, DT, and related concepts

those opportunities in the second part of this paper. Thus, this paper aims at identifying and systematically presenting the taxonomies from the domains that mainly contribute to TAS domain, and at detailing the fundamental and promising research areas that have emerged to enable effective design and operation of TAS in real-world applications. In particular, we consider dependable and resilient computing, cybersecurity and cyber-resilience, trustworthy AI and safe autonomy, as the primary concepts – or "building blocks" – covering the main aspects of interest when dealing with emerging intelligent and adaptive systems operating in critical applications and environments, as represented by TAS.

For the sake of clarity, the **main contributions** of the paper are as follows. In order to highlight the connections of concepts related to dependable and resilient computing, as well as to cybersecurity and cyber-resilience, with TAS properties and technologies, with a focus on classification of changes, evolution, and assessability, we provide a novel structured representation of those taxonomies by using class diagrams (*first contribution*). We address and discuss the most reputable taxonomies developed by diverse scientific communities, in connection with the emerging concepts of trustworthy artificial intelligence. We show how those communities addressed similar yet separate concerns, with many substantial but few formal overlaps in reference concepts and taxonomies; therefore, we try to bridge those differences and provide connections between the most relevant concepts and terminologies (*second contribution*). Finally, we investigate main challenges and promising research directions associated with TAS, based on recent developments within working groups addressing guidelines and regulations for trustworthy autonomy (*third contribution*).

The rest of this paper is structured as follows: Section 2 provides related

works on taxonomies and perspectives that are relevant for TAS, ranging from resilient computing to trustworthy autonomy; Section 3 provides a brief description, structured representation and interconnection of fundamental and emerging concepts in those relevant areas; Section 4 focuses on challenges and promising research directions within TAS as complex, autonomous and adaptive systems; finally, Section 5 draws conclusions.

## 2 Related Works

In this section we mention some works contributing to the definition of taxonomies and research trends within TAS, and more specifically about resilient computing, cybersecurity and trustworthy autonomy. To the best of our knowledge, no work exists providing and interconnecting the relevant concepts and perspectives by using semantic diagrams or structured representations such as class diagrams as we do in this paper. Since we mention related works in the specific subsections of this paper, in this section we only discuss a selection of sources featuring a higher level of generality.

In addition to the seminal papers representing cornerstones in computer dependability and resilience taxonomies (i.e., references [6, 7]), there have been several recent attempts to provide definitions, surveys and reviews about security, resilience and trust, mainly focusing on CPS and IoT (see, e.g., references [8, 9, 10]). A survey paper that is oriented to analyze the different definitions and measures of system resilience is represented by reference [11].

Some related works such as reference [12] study the issue of resilience from algebraic-theoretical perspectives, also relating to concepts such as anti-fragility [13]. In [14], the authors start from the need highlighted in reference [15], and propose a comprehensive metamodel, expressed in GraphQL, which captures the different aspects (from technical to organizational) of CPS resilience. The main goal of this work is to adapt an assessed approach as Systems-Theoretic Accident Model and Processes (STAMP) [16], making connections between classical safety and security concepts.

Within SoS and Critical Infrastructure (CI), resilience is seen as a crucial aspect to deal with emergent threats related to the interrelationships and interdependencies among the different systems and infrastructures. Two references are reported as examples of definition of domain models and metamodels. In reference [4], a comprehensive cyber-physical SoS metamodel and a related System Modelling Language (SysML) profile have been proposed in the framework of the European Union (EU) FP7 project Architecture for Multi-criticality Agile Dependable Evolutionary Open System-of-Systems (AMADEOS). Different aspects that are close to the scope of this paper are considered: evolution in time, behavior emergence, and system interdependencies.

In [17], Security Analysis and Modelling (SecAM) is introduced as a Unified Modelling Language (UML) profile for the modelling and quantitative analysis of CI. Such a profile represents a convergence between dependability concepts, derived from the Modeling and Analysis of Real-time and Embedded systems -

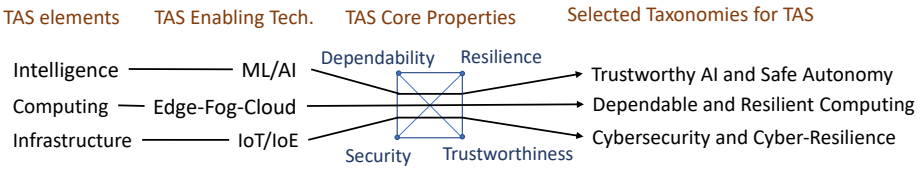


Figure 2: The approach for selecting relevant taxonomies for TAS

Dependability Analysis and Modelling (MARTE-DAM) UML profile [18], and cybersecurity.

The scientific community manifested the need for novel holistic approaches to guarantee the resilience and trustworthiness of increasingly smart, complex and autonomous CPS [15, 19]; however, many basic obstacles need to be tackled, including a common understanding of reference taxonomies and promising directions, which is a gap we aim to fill with this paper.

# 3 Relevant taxonomies for TAS

In order to mature and evolve, a technical domain needs an aggregation of an appropriate scientific community based on a shared and unambiguous language allowing the exchange of knowledge and experiences between its members. The first step in structuring such as language is the definition of a taxonomy, i.e., a scientific classification of objects, subjects, and concepts into possibly hierarchical groups, types and categories, within a certain area of interest. Therefore, in this section, we identify and discuss relevant taxonomies related to the TAS domain. We identified four core concepts driving TAS taxonomy: Trustworthiness, Security, Resilience and Dependability. Such core concepts have guided the selection of taxonomic sub-areas in connection with TAS enabling technologies as depicted in Fig. 2.

These taxonomies have been developed by different research communities, with limited overlaps with each other; particularly, they are: (i) dependable and resilient computing; (ii) cybersecurity and cyber-resilience; and (iii) trustworthy AI and safe autonomy.

## 3.1 Dependable and Resilient Computing

Since its publication in 2004, the taxonomy of dependable and secure computing [6], with its thousands of citations, has defined a de facto standard terminology in the research community of safety-critical, real-time, embedded and fault-tolerant computer systems. Dependability was presented as a complex integrative concept, shortly defined as "*the delivery of service that can justifiably be trusted*" [6]. In the same year, another seminal paper was published about model-based evaluation of dependable and secure computer systems [20]. Four years after, Jean-Claude Laprie, at the time director of research at LAAS-

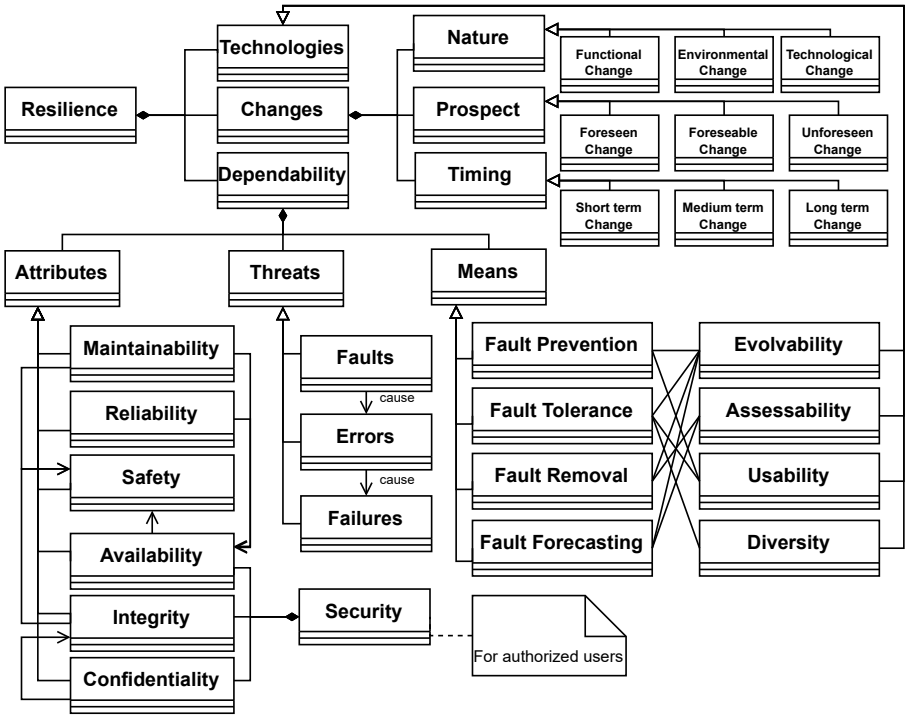


Figure 3: Class diagram for resilient computing and dependability

CNRS (France), published an extension of the dependability taxonomy to account for changes, which the author summarized with the word *resilience* [7], and is nowadays essential to address dependability in intelligent, adaptive and evolving computer systems and environments.

This is the reference taxonomy that is most commonly used by the research community working on safety-critical, real-time, embedded, and fault-tolerant computer systems. A comprehensive description of this taxonomy is reported on references [6], [21], and [7]. For the sake of brevity, we just recall here that dependability is an integrated concept putting together attributes (i.e., maintainability, reliability, availability, safety, confidentiality, and integrity), threats (i.e., faults, errors, and failures), and means (prevention, tolerance, removal, and forecasting). It is worth mentioning that according to this taxonomy, faults are the causes of errors, which are alterations of system state that can generate failures, i.e., effects on system interface. Such a propagation can be subject to different latency. Faults can be classified according to several parameters, including origin (internal vs external), intention (random vs deliberate), etc. Computer security can be seen as a sub-area within computer dependability where the attributes of interest are availability, integrity and confidentiality, while classes of failures are mainly human-made and deliberately malicious.

Resilience is defined in work [7] as the persistence of dependability when facing functional, environmental, or technological changes. In Fig. 3, we have sketched a class diagram providing a compact yet meaningful representation of resilience in relation to dependability, which is expanded into its three main pillars through a composition relationship; changes, which are classified according to their nature, prospect and timing; and technologies, specialized in evolvability, assessability, usability, and diversity, which are put in formal relation with corresponding dependability means. We have also related dependability attributes that were dependent from each other. Such a formal representation, which is missing from the original taxonomy papers, might be further expanded and also detailed with additional interrelationships. However, we prefer to keep the present level of detail that provides good readability and allows defining parallelisms and interconnections with concepts derived from other taxonomies described in the following sections.

While all concepts of dependable and resilient computing are essential to build TAS, some of them are particularly tailored to cope with adaptation to possibly unforeseen environmental changes. We will see how resilience technologies also play an essential role within TAS due to the support for: Evolution – i.e., long-term adaptation; assessability – including explainability; and usability – including interpretability, which is especially important when humans are in the loop to supervise TAS operation such as in semi-autonomous applications or in fall-back operating modes. We will come back to those connections to TAS when addressing specific taxonomies for trustworthy AI and safe autonomy.

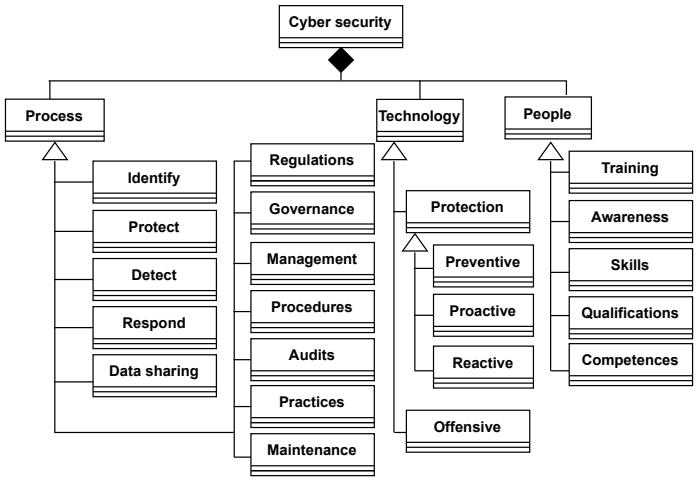


Figure 4: Class diagram for cybersecurity

## 3.2 Cybersecurity and Cyber-Resilience

The concepts of cybersecurity and cyber-resilience enable the creation of trustworthy hyper-connected smart and autonomous systems [22]. The taxonomy outlined in Fig. 4 highlights three cybersecurity pillars that any organization designing, developing, or operating TAS should consider: (i) *Process-related aspects* such as regulations, procedures, etc., to manage threats [23] [24], also involving governance frameworks to guide the regulatory and ethical principles [25], best practices, certification, audit and accountability, security engineering, and trust and privacy principles [26, 27]); (ii) *security technologies* (ranging from offensive measures to defensive ones, whether proactive, preventive and reactive) with application in hardware - HW, software - SW, and network); and (iii) *people* (training and awareness, professional skills and qualifications, etc.).

Those three pillars are essential in new operational ecosystems to dynamically cope with security attacks caused by HW/SW failures possibly coming from the supply chain [28]. In the context of TAS, multiple attacks at SW, HW and network level may arise as detailed in [22], but also as consequences of human-machine interaction — corresponding to point (iii) mentioned above. Any human-made fault generates errors that can lead to security breaches in TAS, thus creating new cyber and physical risks that may impact continuity of operation [29]. Under these circumstances, cybersecurity is a priority condition for deploying TAS in critical operating environments. Because of the complexity of the TAS domain, it is also necessary to consider an extended taxonomy of cybersecurity that includes the holistic concept of "cyber-resilience". In fact, while cybersecurity focuses on attack avoidance and immediate response actions against known risks and vulnerabilities, cyber-resilience extends the focus on strategies and policies to sustain a continuous adaptation to deliver accept-

able operational level against unwanted changes and unexpected environmental conditions.

Due to their complexity and heterogeneity, TAS present an increased surface exposed to cyber-threats, suffer from the under-specified nature of their operation [30], and are prone to the risks of discovered zero-day (i.e., previously unknown and immediately exploitable) vulnerabilities, classified as "unknown unknowns" [31].

Several definitions of cyber-resilience exist. According to reference [32], cyber-resilience refers to the system "ability to continuously deliver the intended outcome despite adverse cyber-events". Such a definition is similar to the one provided by Laprie [7], although the focus is more general, on any type of outcome and entities, especially complex and hierarchically structured organizations, as well as SoS. In reference [33], cyber-resilience is considered in the context of complex systems including cognitive and social domains.

Attempts towards a standardized definition and related taxonomy have come out only recently from NIST [34], hence sometimes terms such as robustness, business continuity and antifragility, used in different domains, have been considered as synonyms although actually representing different meanings. For example, in the Institute of Electrical and Electronics Engineers (IEEE) Standard 610.12.1990, "*robustness is defined as the degree to which a system operates correctly in the presence of exceptional inputs or stressful environmental conditions*". Likewise, business continuity (International Organization for Standardization (ISO) 22301:2019) is process-centric, and it is mostly related to a set of rules and procedures driven by the results of risk assessment to properly react in case pre-identified critical scenarios [35, 36].

Finally, in Taleb's work [13], the concept of antifragility presents strong similarities with the one of resilience. Antifragility is associated with bi-modal risk strategy called "The Barbell", which manifests itself as a good balance between: (i) Strong and weak interactions in network topology; (ii) adaptability and robustness (criticality); and (iii) ascendancy and overhead. Moreover, the work presented in reference [37] mentions that systems should be designed to be antifragile, in the sense that the system has to learn from its experience, adapt to unforeseen events, and grow stronger in the face of adversity. This is especially relevant for TAS due to their ability to learn an adapt through AI and ML.

Regardless of those different definitions, a convergence in a taxonomy for TAS can be foreseen regarding the aspects of bounce-forward (or Building Back Better, BBB), learning for "evolvability" and the emergence of a resilient behavior as a result of a continuous adjustment to seek a dynamic equilibrium in the system [38]. In TAS, such a continuous adjustment is expected to be performed by leveraging on AI and ML, which are therefore enablers for advanced cyber-resilience.

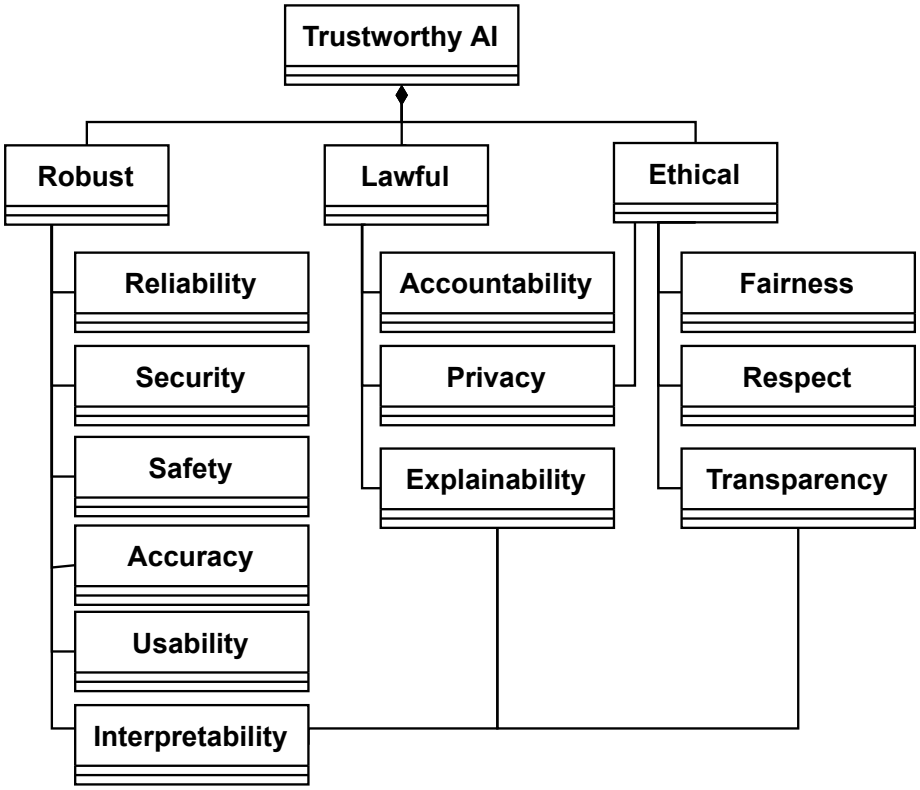


Figure 5: Class diagram for trustworthy AI

### 3.3 Trustworthy AI and Safe Autonomy

In this section, we address the taxonomy of "Trustworthy AI", also named with slightly different terms such as safe/trusted autonomy, which has been associated a different and specific meaning compared to the trustworthiness mentioned in [6]. Such a taxonomy has been recently developed by a community of AI experts dealing with threats and challenges of autonomous robots and vehicles, including ethical dilemmas and legal issues such as accountability in case of accidents. Most of the resilience-related issues addressed by those experts have been summarized with the term "robustness" [39], which seems to be a synthesis of attributes such as reliability, safety and security, plus fault-tolerance, although no formal or structured definition exists of robustness in the context of trustworthy AI, especially in connection with learning, adaptation and evolution capabilities. Although robustness and resilience may sound similar, the former is more focused on managing input data perturbation, and does not necessarily include all the resilience aspects we have addressed in the previous sections. Another common term in the field of trustworthy AI is "explainability", which refers to the possibility of explaining the inner behavior of intelligent systems, although ML might behave as a "black-box" compared to traditional control algorithms and engineering models. Such an issue with unpredictability of machine learning systems is sometimes also referred to as "opacity". Explainability is not an attribute or property defined within the other taxonomies introduced in the previous sections, however it can be associated with "assessability" [7] that is in turn connected with the dependability means of "fault-removal" and "fault-forecasting" (see Fig. 3): having an AI which can be explained allows to assess models, predict failures at run-time before they can generate severe consequences, and also allow forensic investigations in case of accidents.

To be trusted, AI should also be "ethical" [40], which is an attribute addressing aspects such as non-discrimination and proportionality, which is difficult to associate to any of the cyber-resilience concepts introduced in the previous sections. A connection can be made with the "confidentiality" attribute of AI if the focus is on non collecting and disclosing any private information and sensitive user data if not strictly required. However, all the aspects related to possible ethical bias and dilemmas when autonomous systems must take important decisions are not addressed in traditional cyber-resilience taxonomies that were not focused on managing high-level intelligent behaviors and full autonomy.

In Fig. 5, we summarize the three main characteristics of trustworthy AI that are essential within TAS. We connect robustness to reliability, safety, and security, the latter including both cyber- and physical security; usability (which is an attribute also introduced in the computer resilience taxonomy), needed to consider human-factors, social implications, and ergonomics, as indicated by relevant expert groups [39]; accuracy, which is a fundamental trustworthiness characteristic of intelligent classifiers when taking decisions supporting or replacing humans; and interpretability, which is essential for the humans to understand, trust, and validate automatic decisions. We connect legality to accountability, explainability (which might be essential to obtain safety certification, as

discussed above), and privacy, which is intended here as confidentiality from a legal perspective. We connect ethical aspects to fairness (integrating equity, equality, non-discrimination, absence of bias, inclusiveness, etc.), respect (for diversity, freedom, etc.), and transparency in providing evidence of the aforementioned ethical attributes such as non-discrimination. We made an effort to limit redundancies in such classification, considering that several characteristics have similar meanings. However, note that there are several meaningful overlaps and interconnections between those concepts: For instance, privacy — in addition to be a legal requirement — can be classified in relation with its ethical dimension and the more general attribute of respect, i.e., respect for privacy; interpretability, explainability, and transparency, as explained above, are clearly interrelated, although they represent different views on the same issue, respectively, from robustness, legal, and ethical perspectives.

# 4 TAS Challenges and Research Directions

In this section, we identify relevant challenges in the TAS domain, as well as the related opportunities and promising research directions. Those research directions have been mainly derived from the IEEE Approved Draft about "*Standard for Transparency of Autonomous Systems*" [41].

As discussed in the previous section, the lack of widely accepted guidelines and regulations has been a big hurdle to the development of TAS. Since traditional standards for critical systems cannot be easily applied to intelligent systems due to their limitations in transparency and predictability, development of new standards – which is a challenge itself – is essential to define key challenges and research directions depending on real-world objectives and non-functional requirements. In such a context, it is worth mentioning the IEEE Approved Draft "*Standard for Transparency of Autonomous Systems*" [41], which stresses the concept of transparency as the prime driver in achieving trust from final user, as well as from system assessor during validation activities. The standard defines different levels of transparency depending on involved stakeholders and also proposes a System Transparency Assessment (STA) and a System Transparency Specification (STS).

The theme of transparency is also central in DIN SPEC 92001-2:2020-12 [42], although applied to the development process rather than on intelligent applications. Robustness is considered as a primary mean to achieve trustworthiness and six steps are defined, including the one identified with "define AI malfunction per automated task". Likewise, a recent deliverable of the SCSC Safety of Autonomous Systems Working Group (SASWG) is the third release of the "*Safety Assurance Objectives for Autonomous Systems*" [43]. This comprehensive work addresses the problem of confidence in Trustworthy Autonomous Systems (TAS) within three different frameworks of growing levels of abstraction and complexity:

- *Computation level*, addressing implementation at the software and hardware levels (associated to fault prevention);

- *Autonomy architecture level*, addressing how computations can be integrated into a system, or platform including fault-tolerance;
- *Platform level*, addressing what the TAS should (and should not) do, and related effects on the operational environment (requirements engineering).

Such a document represents an effective synthesis of the traditional safety-critical principles and the challenges originated by AI technologies. For all three mentioned levels, the SASWG defines different objectives, including the relevant ones reported in Table 1, where the first column is related to future research directions mentioned in this paper.

## 4.1 Mitigating Threats to TAS

**Challenges:** Mitigating emerging threats is currently one of the most crucial challenges within TAS. All countries, including EU member states, showed an increasing interest in the possible implications and misuse of AI in critical applications, in terms of safety, data privacy, and homeland security. In 2021, a task force of the Centre for European Policy Studies (CEPS) published a report addressing the most important cyber-threats to AI systems [44]. Other valuable documents have been published by the European Union Agency for Cybersecurity (ENISA) in 2021 [45], and by the European Telecommunications Standards Institute (ETSI) [46] in 2020.

The greater vulnerability to cyber-attacks is due to the larger attack surface of ML systems, which can suffer from both traditional HW/SW threats and new threats related to the training processes, the interaction with the external environments (i.e., sensing and actuation), and system adaptation and evolution at run time. Emerging AI threats, which may be launched in different scales in TAS depending on the adversary's ability to gain access to data and models (black, gray and white box) [47], include but are not limited to:

- *Input attacks/evasion*, where input to the ML systems are intercepted and changed — in a perceivable or unperceivable manner — into those data patterns in order to generate a failure;
- *Poisoning attacks*, where training datasets, learning algorithms, or models are "poisoned" — i.e., corrupted — to compromise learning;
- *Backdoor attacks*, where attackers insert some sort of "smart backdoor" (also employing special data patterns) to be exploited at run-time;
- *Reverse engineering*, aimed at extracting input-output pairs from ML models.

A comprehensive survey of those attacks can be found in reference [48], while an ENISA technical report frames the attacks described above into a framework addressing AI threats. Such framework embraces different phases of system life-cycle and includes several non-technical aspects as legal threats against ML systems.

Table 1: SASWG objectives.

| **TAS Future research direction** | **SASWG objective** |
|---|---|
| Threat Mitigation | COM1-4 Adverse effects arising from distribution shift are protected against<br>COM3-2 Typical errors are identified and protected against<br>ARC1-4 Adversarial attempts to disrupt the computation are tolerated |
| Explainability of autonomous decisions | COM3-3 The algorithm's behaviour is explainable<br>ARC2-1 Relevant information is presented to interacting parties<br>PLT3-2 Suitable interfaces are provided for people that may interact with the platform |
| Resilience assessment | COM2-2 Non-functional requirements imposed on the algorithm are defined and satisfied.<br>COM2-3 Algorithm performance is measured objectively<br>COM4-1 The software is developed and maintained using appropriate standards<br>PLT1-6 Operational monitoring is sufficient to identify and support the mitigation of new hazards, including emerging cybersecurity threats. |
| Intelligent cybersecurity | ARC1-4 Adversarial attempts to disrupt the computation are tolerated<br>PLT1-4 The specified behaviour is safe in the presence of faults and failures, as well as foreseeable misuse and abuse. |
| Digital Twins | COM2-6 The test environment is appropriate<br>ARC3-1 Inappropriate or unauthorised adaptations do not occur<br>PLT1-5 The behaviour of the platform is verified. |
| Self-Healing | ARC1-1 Failures of sub-systems that provide computation inputs are tolerated<br>ARC1-2 Operational inputs inconsistent with training inputs are tolerated<br>ARC1-3 Faults and failures internal to the computation are tolerated |

**Research perspectives:** As shown above, there are some efforts to exploit AI threats, but not so much to mitigate them. In reference [49], the authors identify some security challenges for autonomous systems and robots, where security-by-design, access control, and automation for prevention (as discussed below) are current priority research lines. Also, the survey on AI threats and countermeasures detailed in [47] points out the relevance of protection in software- and hardware-based data collection, malware control, and the relevance of understanding the behavior of each attack to derive defensive measures. Thus, the survey identifies for each AI attack possible mitigation strategies (e.g., data poisoning could be mitigated by data sanitization), and lists a set of defense methods for ML models and data.

## 4.2 Explainability of autonomous decision in TAS

**Challenges:** The explainability of autonomous decisions taken by the embedded AI is a fundamental feature that AS have to exhibit for the sake of trustworthiness. However, eXplainable AI (XAI) is one of the paramount open challenges within TAS. XAI aims to mitigate issues related to predictability of intelligent systems, and to manage the problem of avoiding "super-human" agents, as discussed in reference [50]. Recent advancements in this field are summarized by state-of-art papers such as work [51], which investigates the issue from a general perspective, and work [52], which focuses on medical XAI. From those references, we highlight the main challenges listed below.

- Clear definition of explainability: The literature reports different XAI approaches, and there are only a few attempts to provide a unifying vision [53]. The problem is mainly related to the definition of a set of metrics and performance indices related to explainability.

- Interpretability confidence: To improve trustworthiness and foster the adoption of XAI approaches, solutions that can provide a possibly quantitative estimation of the likelihood and the uncertainty of such explanation should be pursued. Papers addressing such problem are listed as references [54] and [55], where Bayesian Networks are used as probabilistic ML models.

- Trade-offs between interpretability and performance: In reference [51], a general law of balance between is proposed, showing that ML models that perform better (e.g., DL) are less explainable.

- Confidentiality: As AI systems are developed by companies with large financial investments, there should be an equilibrium between how much such algorithms should be transparent to a final user and the protection of the Intellectual Property, since adversarial networks can also infer information by reverse engineering (see Subsection 4.1).

- Standardisation: There is a necessity to define standards and guidelines to certify explainable systems. One attempt in this direction is constituted by

the XAI standard under development by the Computational Intelligence Society Standards Committee (CIS/SC) Working Group of the IEEE [56].

- Legal and ethical aspects: As a fundamental objective of XAI is to increase the level of trust in autonomous systems, one paramount goal is to understand the boundaries and constraints under which explanation of intelligent machines could be used in a legal lawsuit. The discussion of AI ethics is also open and very much debated [57].

**Research perspectives:** The research community working on XAI has defined some techniques and practical frameworks to develop interpretable models and explanations of ML decisions that can be comprehensible to humans. Such solutions are framed into the two main approaches of XAI: model-agnostic, which do not require knowledge about the internal structure of the ML algorithm (i.e., black box), and model-specific, which require knowledge about how the algorithms are structured (i.e., white/grey box). Most widespread tools are both model-agnostic and model-specific, such as LIME[1], SHAP[2], and ELI5[3].

### 4.3 Resilience Assessment of TAS

**Challenges:** In AS, self-awareness [58] allows systems to dynamically adapt to changing situations in order to survive to external stressors. To that aim, the holistic estimation and quantification of resilience in TAS through data-driven appropriate metrics is necessary and crucial to enable specific warnings and response actions. Dependability attributes such as reliability and safety can be used to provide quantitative indicators of certain system properties; however, they focus on specific parameters and are unable to capture the appropriate level of complexity of the system needed for an effective adaptation to changing and unwanted conditions. In fact, basic dependability attributes do not take into account all resilience and performability-related aspects, such as response times, time to reach the expected dynamic equilibrium after a disruption, or operational capacity after reconfiguration. Unfortunately, no single widely accepted metric exists for a consistent, quantitative measurement of TAS resilience, but several have been proposed in the scientific literature [59]. For example, TAS data-driven resilience measurement can be:

- Direct, based on time-dependent performance assessment of critical functionalities [60], where the functionality of a system is defined as a non-stationary stochastic process and each ensemble is a piece-wise continuous function [61] [62]. This approach requires online computation, sufficient data availability and completeness, as well as a very detailed system specification and modeling knowledge. Direct resilience assessment metrics include measures of: time, e.g., when performance degrades below a threshold, and when it is restored above the threshold; performance, e.g., the

[1] https://homes.cs.washington.edu/ marcotcr/blog/lime/
[2] https://github.com/slundberg/shap
[3] https://eli5.readthedocs.io/

area under the performance curve from the time when performance degradation starts and the time when recovery is complete; resources, available and consumed for adaptation and recovery; and impact, e.g., number of users affected [63].

- Indirect, by analyzing only the potential for resilience represented by system capabilities [64]. For instance, in reference [65] the Functional Resonance Analysis Method has been adopted to estimate the System Resilience Index (SRI), a proxy indicator to assess system resilience before a critical event happens. It is computed by analyzing the four resilience capabilities (anticipate, respond, monitor, and learn) available in the system at a certain instant of time by using a data-driven approach.

Unfortunately, each of those approaches presents limitations and provide a partial understanding of the system with the risk of not taking well-informed decisions.

**Research perspectives:** One of the most promising research line able to overcome the limitations of the current approaches is represented by the attempt of merging them into a single holistic view where resilience is considered as a property that emerges from an interaction dynamics between adaptive capacity and coping ability of a system as elaborated in references [38] and [66] It is necessary to reconcile concepts and terminologies among the different domains involved in TAS to define an cross-domain and wider accepted ontology able to drive the development of metrics for resilience.

Moreover, all those measures should be put in relation with a service level acceptance criteria. Namely, a set of constraints and relationships enabling identification of the subset of system state space consisting of all those states where the service delivered can be considered correct and acceptable by the users [67]. Acceptability can be regulated by agencies and reference standards [68], or by market/customer demand through Service Level Agreements (SLA), but when operations are supported by multiple components interacting each other in a multi-stakeholder scenario, the definition of service level acceptance becomes blurred and subjective; that prevents the adoption of the autonomous evidence-driven decision making mechanisms required for TAS.

In summary, in TAS domain we identify the need for defining:

- Unified holistic models of cyber-resilience, accommodating different attributes and metrics, combining direct and indirect resilience assessment methods.

- Novel and standardised sets of cost-effective quali-quantitative metrics to allow a cyber-resilience self-assessment, possibly merging data analytics and expert judgment combining together data-driven and model-driven approaches.

- Relations between service level acceptance and service disruption impact, which is defined as a set of objective impact metrics (e.g., the extent of

the network impacted in terms of users) that can be used as a computable fast-forward proxy indicators for resilience.

- Transparent, consensus-driven and reliable methods to select the pertinent indicators from relevant standards and guidelines, and to extend the selected set with specific resilience indicators proposed by experts.

### 4.4 Intelligent cybersecurity for TAS

**Challenges:** As discussed in Section III.B, cybersecurity is essential for TAS in order to provide optimal and increased protection against potential threats [22], such as stealthy attacks, malware or zero-day HW/SW attacks. AI technologies for cybersecurity can be used to dynamically improve detection and response processes. According to the European Union Agency for cybersecurity (ENISA) [69], there are three main dimensions of application: (i) AI to support cybersecurity, (ii) attacks to AI technology, and (iii) cybersecurity for AI.

The former dimension is one of the most widespread research areas in the current literature [70, 71], which focuses on improving cybersecurity by applying advanced techniques capable of detecting, locating and neutralizing potential threats. Reference [72] explicitly highlights the current trends to design advanced ML intrusion detection systems in the new industrial ecosystems; and work [49] in TAS operational environments, detailing AI opportunities for defense automation at multiple levels. Anomaly detection approaches based on ML techniques are also emerging to identify unexpected fluctuations in nominal behavior of complex systems due to unknown threats of random or malicious origin [73]. There exist multiple studies in the literature exploring the role of ML for intrusion and anomaly detection, contemplating various techniques such as traditional ML models, DL, federated learning or collaborative detection [72, 74, 75, 76, 77]. The use of these techniques should be limited to a phase of training, learning and continuous readjustment to reduce the false positive rate as proposed in [78], where the goal is to provide collaborative strategies for retraining of ML models. Other works consider AI and ML to support Security Information and Event Management (SIEM) operations [79] used at Security Operation Centers (SOC); cyber-threat intelligence [80]; situational awareness [71, 81]; self-testing [44]; and self-healing [70]. We will address the latter in Section 4.6.

There are, however, certain limitations when adapting automated techniques for cybersecurity. According to references [70] and [71], the use of AI leads to the need to rely on highly powerful HW/SW resources to manage large volumes of heterogeneous data, which in turn forces the usage of specific big data techniques to clean noise and duplication, and manage and normalize multi-source data. This data management and the inherent complexities of ML models may further limit the system's *rapid response* to anomalous events. In such a process, the reaction can consist from simple notifications to automated responses based on intelligent and collaborative methods supported by complex risk assessment processes [82]. The work in reference [71] highlights the lack of *transparency in*

*decision-making* processes, which makes it difficult to trust automated decisions in critical contexts.

Regarding the second dimension, ENISA highlights the weakness of ML models in the face of various adverse scenarios. In [69], a taxonomy of attacks is presented, ranging from attacks related to data interference, data theft and model disclosure, to attacks related to sample and model manipulation, illicit access and analysis, privacy leakage during data operations, among other kinds of attacks; as also stated in Section 4.1. For that reason, there is still a need for addressing those accesses that can violate *data privacy and integrity* of samples and models [71] to generate high rates of false positives and negatives.

**Research perspectives:** According to the challenges identified, the emerging research lines are focused on adjusting ML models to maximize detection *accuracy* [71, 78]. In addition, research on robustness and protection (corresponding to the third dimension given by ENISA) are still necessary to avoid illegitimate accesses or exploits that can impact on automation and decision-making, as well as on detection and prediction accuracy to get correct analysis and responses against potential threats. Experts also need new models that are *resilient to noisy and incomplete data*, *flexible and adaptable* to unforeseen changes, and *collaborative* (e.g., agent-based intelligent systems) to avoid inconsistencies [44].

## 4.5 Digital Twins for TAS

**Challenges:** DT have emerged as a valuable paradigm for online predictive analytics and prognostics in TAS, e.g., for optimizing safety and security in self-driving vehicles and autonomous robots, as shown in references [83, 84, 85]. These approaches are mainly powered by AI and ML [86] to create predictive models for run-time analysis and proactive response, in order to automatically anticipate and respond to errors and failures, possibly including safety hazards. In fact, the application of formal verification methods to TAS is a well-known challenge due to complexity and unpredictability[4], therefore detailed run-time models based on simulation can help cope with online anomaly detection and management, e.g., by implementing safety-cages/envelopes [87].

Experts have studied the effectiveness of DT technology to support cyber-resilience in critical contexts there TAS can be required to operate. For example, in [88] DT are used for incident prevention and response to maximize cyber-resilience in the context of power grid. In [89], the authors propose DT in the context of smart-cities to model spatial, logical and temporal interdependencies in urban environments modelled as multi-layered CPS. Policies for obtaining a resilient behavior from the system in face of changing conditions are calculated by simulating multiple scenarios based on DT. In [2], a conceptual DT framework is presented that can be applied to monitor and improve resilience of autonomous CPS, supported by edge-fog-cloud computing. In reference [90],

[4]Shonan Meeting on "Formal Methods for Trustworthy AI-based Autonomous Systems", https://shonan.nii.ac.jp/seminars/178/

the authors present the benefits of simulation to manage crisis situations by analyzing how technology can impact the different resilience phases, such as anticipation, monitoring, response and learning. The works in references [91] and [92] focus on DT for online cyber-defense, and the challenges that the technology can bring to such a application area. A set of cybersecurity use cases are described in reference [92], such as anomaly and intrusion detection, event management to improve situation awareness, testing, training, privacy and legal compliance.

**Research perspectives:** In order to fully leverage on DT potential and create TAS-based operating environments, some relevant aspects must be considered as research directions, such as *accuracy*, *usability*, *expandability*, and *security* [93]. In this case, accuracy refers to how digital models must characterize the specific configuration, topology, traffic load and dynamics of the physical system with sufficient accuracy to reproduce the behavior of the physical system. Usability is associated with the ease of creation, maintenance and final use (e.g., for operators/maintainers attending emergency scenarios), bearing in mind the degree of data computation and interpretation within DT, as well as its heterogeneity. Expandability is relevant to ensure effectiveness in industrial contexts. DT must scale up easily to simulate increasingly complex CPS. As for cybersecurity, the objective is protect DT themselves, as digital models might be integrated using insecure architectures. This is detailed in reference [93], where a comprehensive taxonomy of DT threats along with security measures can be found. Some of these measures are: (i) access control to protect critical resources and decision making, since DT can be targets for attackers interested in sabotage or in getting valuable information about physical components; (iii) advanced intrusion detection; (iv) cryptography; and (v) appropriate auditing measures, especially when DT is applied for emergency response and crisis management.

## 4.6 Self-Healing TAS

**Challenges:** Self-healing is not a completely new concept, since it has been mentioned even in the seminal dependability taxonomies mentioned at the beginning of this paper, as an extension of — or rather a perspective over — fault-tolerance, in order to achieve resilience and plasticity. According to recent surveys on the subject, first adoptions of self-healing — a specialization of self-management/self-adaptation — can be associated with Defense Advanced Research Projects Agency (DARPA) projects and with the concept of autonomic computing. Recent worth mentioning papers addressing self-healing in CPS and IoT are: reference [94], where self-healing is embedded into "smart-troubleshooting" to improve resilience of interconnected and heterogeneous devices in a holistic system-of-systems perspective, including pragmatic aspects of human-based information processing and security management; and reference [95], where several relevant self-* mechanisms have been addressed.

The challenges related to self-healing for TAS can be addressed at the following levels.

- Software. In self-healing software, fixes are applied to solve run-time problems. Challenges include dynamic generation, verification of correctness (and related testing), and evaluation of non-functional properties [96].

- Computing. . The open challenges identified at this level are as follows: state-flapping, an emergent behaviour related to the presence of oscillations occurring between states (and related actions) with and impact on optimal system operation; benchmarking, i.e., how the performance of the system can be evaluated, focusing on the evaluation of different and contrasting features in few global indices; interoperability, i.e., how different autonomic systems interact among them [97].

- Critical infrastructures. At this level, the implementation of self-healing is extremely challenging in terms of: required redundancy, i.e., the presence of up-to-date replicas of running components to duplicate functionalities; coordination, due to synchronization problems in concurrent systems; and self-stabilization, i.e., the capability to dynamically control transient faults by reaching a safe state into a finite number of steps [26].

As a general consideration, as the size of the system increases (e.g., from software to CI), it is hard to define and assess effective self-healing mechanisms. As realistic test benches are not possible for CI, DT might represent a valid approach to support self-healing (see Subsection 4.5) [2].

**Research perspectives:** On the base of the challenges identified above, future research directions in self-healing TAS include improvement of off-line design methods for maintainability, as well as definition of proper metrics to trigger healing procedures (benchmark problem). The Monitor, Analyse, Plan, Execute, Knowledge (MAPE-K) control loop [98] is also expected to evolve as a general framework to develop self-* systems in order to address computing and critical infrastructure challenges such as state-flapping and self-stabilization. Another planned advancement focuses on ML for smarter threat management (i.e., intelligent detection, diagnosis, and reconfiguration/response) in order to enable sophisticated and holistic approaches to fault-tolerance and resilience, including those based on on-line conformance checking [99].

# 5 Conclusions

In this paper, we have provided a compendium of terms and concepts to address relevant aspects within TAS. In addition, we used structured representations of related taxonomies to highlight the crucial concepts connected to TAS, especially regarding intelligent and adaptive behaviors in presence of changes and evolution. We have also presented the most challenging research directions to cope with predictability and assessability in presence of opacity and uncertainties; those properties are vital, as well as essential to support trust and certification of those complex and heterogeneous systems against international standards and regulations. Moreover, in this study we have shown that the

paradigm of trustworthy autonomy, which applies to intelligent systems operating in critical applications, includes and extends the concepts addressed in existing taxonomies, such as the ones addressing computer dependability and cyber-resilience. We made an effort towards providing a more systematic presentation and mapping of those concepts, and also addressed some of the main research challenges that must be tackled to increase trust in autonomous CPS in presence of rapidly evolving functionalities, new environments, uncertainties, emerging threats, and "unknown unknowns".

We believe that proper knowledge and awareness of the concepts, perspectives and opportunities presented in this paper is a prerequisite for a further formalization into comprehensive TAS domain representations, based on ontologies, specification languages, and semantic models, to support model-based/model-driven engineering at all phases of TAS life-cycle, and possibly enable paradigms such as trustworthiness-by-design [100].

# References


[1] F. Flammini, *Resilience in CPS*. Springer Berlin Heidelberg, 2019. [Online]. Available: https://doi.org/10.1007/978-3-642-27739-9_1728-1

[2] ——, "Digital twins as run-time predictive models for the resilience of cyber-physical systems: a conceptual framework," *Trans. R. Soc. A*, vol. 379, no. 20200369, 2021.

[3] E. Bellini, P. Ceravolo, and P. Nesi, "Quantify resilience enhancement of uts through exploiting connected community and internet of everything emerging technologies," *ACM Trans. Internet Technol.*, vol. 18, no. 1, oct 2017. [Online]. Available: https://doi.org/10.1145/3137572

[4] A. Bondavalli, S. Bouchenak, and H. Kopetz, *Cyber-Physical Systems of Systems: Foundations – A Conceptual Model and Some Derivations: The AMADEOS Legacy*. Springer International, 2016.

[5] D. J. Hand and S. Khan, "Validating and verifying ai systems," *Patterns*, vol. 1, no. 3, p. 100037, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666389920300428

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

[7] J. Laprie, "From dependability to resilience," in *Proc. IEEE Conf. Dependable Systems and Networks, DSN'08*, 2008.

[8] A. Gawanmeh and A. Alomari, "Taxonomy analysis of security aspects in cyber physical systems applications," in *2018 IEEE Intl Conf. on Communications Workshops (ICC Workshops)*, 2018, pp. 1–6.

[9] F. M. R. Junior and C. A. Kamienski, "A survey on trustworthiness for the internet of things," *IEEE Access*, vol. 9, pp. 42 493–42 514, 2021.

[10] R. Roman, J. Lopez, and S. Gritzalis, "Evolution and trends in the security of the internet of things," *IEEE Computer*, vol. 51, pp. 16–25, 07/2018 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8423133/

[11] S. Hosseini, K. Barker, and J. Ramirez-Marquez, "A review of definitions and measures of system resilience," *Reliability Engineering and System Safety*, vol. 145, pp. 47–61, 2016.

[12] V. De Florio, "Antifragility = elasticity + resilience + machine learning models and algorithms for open system fidelity," *Procedia Computer Science*, vol. 32, pp. 834–841, 2014, the 5th Intel Conf. on Ambient Systems, Networks and Technologies (ANT-2014).

[13] N. N. Taleb and R. Douady, "Mathematical definition, mapping, and detection of (anti)fragility," *Quantitative Finance*, vol. 13, no. 11, pp. 1677–1689, 2013.

[14] G. Bakirtzis, T. Sherburne, S. Adams, B. Horowitz, P. Beling, and C. Fleming, "An ontological metamodel for cyber-physical system safety, security, and resilience coengineering," *SoSYM*, 2021.

[15] G. Bakirtzis, G. Ward, C. Deloglos, C. Elks, B. Horowitz, and C. Fleming, "Fundamental challenges of cyber-physical systems security modeling," 2020, pp. 33–36.

[16] N. Leveson, *Engineering a safer world: systems thinking applied to safety*. MIT press, 2011.

[17] R. Rodríguez, J. Merseguer, and S. Bernardi, "Modelling security of critical infrastructures: A survivability assessment," *Computer Journal*, vol. 58, no. 10, pp. 2313–2327, 2014.

[18] S. Bernardi, J. Merseguer, and D. Petriu, "A dependability profile within marte," *SoSYM*, vol. 10, no. 3, pp. 313–336, 2011.

[19] D. Tokody, J. Papp, L. Iantovics, and F. Flammini, *Complex, Resilient and Smart Systems*. Springer, 2019.

[20] D. Nicol, W. Sanders, and K. Trivedi, "Model-based evaluation: from dependability to security," *IEEE Transactions on Dependable and Secure Computing*, vol. 1, no. 1, pp. 48–65, 2004.

[21] J. Laprie, "Resilience for the scalability of dependability," in *Fourth IEEE International Symposium on Network Computing and Applications*, 2005, pp. 5–6.

[22] F. Jahan, W. Sun, Q. Niyaz, and M. Alam, "Security modeling of autonomous systems: a survey," *ACM Computing Surveys (CSUR)*, vol. 52, no. 5, pp. 1–34, 2019.

[23] NIST, "Framework for Improving Critical Infrastructure Cybersecurity," 2018. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.04162018.pdf

[24] A. Gawanmeh and A. Alomari, "Taxonomy analysis of security aspects in cyber physical systems applications," in *2018 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2018, pp. 1–6.

[25] J. Sifakis, "Can we trust autonomous systems? boundaries and risks," in *Automated Technology for Verification and Analysis*, Y.-F. Chen, C.-H. Cheng, and J. Esparza, Eds. Cham: Springer International Publishing, 2019, pp. 65–78.

[26] C. Alcaraz and S. Zeadally, "Critical infrastructure protection: Requirements and challenges for the 21st century," *International Journal of Critical Infrastructure Protection*, vol. 8, pp. 53–66, 2015.

[27] F. J. R. Lera, C. F. Llamas, Ángel Manuel Guerrero, and V. M. Olivera, "Cybersecurity of robotics and autonomous systems: Privacy and safety," in *Robotics*, G. Dekoulis, Ed. Rijeka: IntechOpen, 2017, ch. 5. [Online]. Available: https://doi.org/10.5772/intechopen.69796

[28] S. Fischer-Hübner, C. Alcaraz, A. Ferreira, C. Fernandez-Gago, J. Lopez, E. Markatos, L. Islami, and M. Akil, "Stakeholder perspectives and requirements on cybersecurity in europe," *Journal of Information Security and Applications*, vol. 61, no. 102916, 09/2021 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2214212621001381

[29] C. Alcaraz and J. Lopez, "Analysis of requirements for critical control systems," *International Journal of Critical Infrastructure Protection (IJCIP)*, vol. 5, p. 137–145, 2012 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1874548212000455

[30] J. R. Wilson, B. Ryan, A. Schock, P. Ferreira, S. Smith, and J. Pitsopoulos, "Understanding safety and production risks in rail engineering planning and protection," *Ergonomics*, vol. 52, no. 7, pp. 774–790, 2009.

[31] J. Park, T. P. Seager, P. S. C. Rao, M. Convertino, and I. Linkov, "Integrating Risk and Resilience Approaches to Catastrophe Management in Engineering Systems: Perspective," *Risk Analysis*, vol. 33, no. 3, pp. 356–367, 2013.

[32] F. Björck, M. Henkel, J. Stirna, and J. Zdravkovic, "Cyber resilience – fundamentals for a definition," in *New Contributions in Information Systems and Technologies*. Springer Intl Publishing, 2015, pp. 311–316.

[33] I. Linkov and A. Kott, *Fundamental Concepts of Cyber Resilience: Introduction and Overview*. Springer, 2018, pp. 1–25.

[34] R. Ron, P. Victoria, G. Richiard, B. Deborach, and R. Mcquaid, "Developing cyber-resilient systems: A system, security engineering approach," 2021.

[35] C. Alcaraz, "Cloud-assisted dynamic resilience for cyber-physical control systems," *IEEE Wireless Communications*, vol. 25, no. 1, pp. 76–82, 02/2018 2018.

[36] C. Alcaraz and S. Wolthusen, "Recovery of structural controllability for control systems," in *8th IFIP WG 11.10 Intl Conf on Critical Infrastructure Protection*, vol. 441. Springer, 2014, pp. 47–63.

[37] K. H. Jones, "Engineering antifragile systems: A change in design philosophy," *Procedia Computer Science*, vol. 32, pp. 870–875, 2014, 5th Intl Conf on Ambient Systems, Networks and Technologies (ANT-2014).

[38] E. Bellini and S. Marrone, "Towards a novel conceptualization of cyber resilience," in *2020 IEEE World Congress on Services (SERVICES)*, 2020, pp. 189–196.

[39] E. Hickman and M. Petrin, "Trustworthy AI and Corporate Governance: The EU's Ethics Guidelines for Trustworthy Artificial Intelligence from a Company Law Perspective," *European Business Organization Law Review*, vol. 22, no. 4, pp. 593–625, 2021.

[40] F. Alaieri and A. Vellino, "Ethical decision making in robots: Autonomy, trust and responsibility," in *Social Robotics*. Springer, 2016, pp. 159–168.

[41] "IEEE Approved Draft Standard for Transparency of Autonomous Systems," *IEEE P7001/D4, October 2021*, 2021.

[42] "Artificial intelligence - life cycle processes and quality requirements - part 2: Robustness," *DIN SPEC 92001-2*, 2020.

[43] The SCSC Safety of Autonomous Systems Working Group, *Safety Assurance Objectives for Autonomous Systems V3*, 2022.

[44] L. Pupillo, S. Fantin, A. Ferreira, and C. Polito, "Artificial Intelligence and Cybersecurity," 2021. [Online]. Available: https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges

[45] ENISA, "Securing Machine Learning Algorithms," 2021. [Online]. Available: https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms/@@download/fullReport

[46] SAI working Group, "ETSI GR SAI 004 V1.1.1 - Securing Artificial Intelligence (SAI)," 2020. [Online]. Available: https://www.etsi.org/deliver/etsi_gr/SAI/001_099/004/01.01.01_60/gr_SAI004v010101p.pdf

[47] Y. Hu, W. Kuang, Z. Qin, K. Li, J. Zhang, Y. Gao, W. Li, and K. Li, "Artificial intelligence security: Threats and countermeasures," *ACM Comput. Surv.*, vol. 55, no. 1, nov 2021. [Online]. Available: https://doi.org/10.1145/3487890

[48] D. Jeong, "Artificial intelligence security threat, crime, and forensics: Taxonomy and open issues," *IEEE Access*, vol. 8, pp. 184 560–184 574, 2020.

[49] H. He, J. Gray, A. Cangelosi, Q. Meng, T. McGinnity, and J. Mehnen, "The challenges and opportunities of artificial intelligence for trustworthy robots and autonomous systems," in *2020 3rd International Conference on Intelligent Robotic and Control Engineering (IRCE)*. IEEE, 2020, pp. 68–74.

[50] S. Kate Devitt, *Trustworthiness of Autonomous Systems*. Springer, 2018, pp. 161–184.

[51] A. Barredo Arrieta and et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[52] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): Toward medical xai," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4793–4813, 2021.

[53] D. Shin, "The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai," *International Journal of Human Computer Studies*, vol. 146, 2021.

[54] B. Mihaljević, C. Bielza, and P. Larrañaga, "Bayesian networks for interpretable machine learning and optimization," *Neurocomputing*, vol. 456, pp. 648–665, 2021.

[55] F. Flammini, S. Marrone, R. Nardone, M. Caporuscio, and M. D'Angelo, "Safety integrity through self-adaptation for multi-sensor event detection: Methodology and case-study," *Future Generation Computer Systems*, vol. 112, pp. 965–981, 2020.

[56] IEEE, "P2976 - standard for XAI – eXplainable Artificial Intelligence - for Achieving Clarity and Interoperability of AI Systems Design." [Online]. Available: https://standards.ieee.org/ieee/2976/10522/

[57] A. Kerr, M. Barry, and J. Kelleher, "Expectations of artificial intelligence and the performativity of ethics: Implications for communication governance," *Big Data and Society*, vol. 7, no. 1, 2020.

[58] N. Dutt, C. S. Regazzoni, B. Rinner, and X. Yao, "Self-awareness for autonomous systems," *Proceedings of the IEEE*, vol. 108, no. 7, pp. 971–975, 2020.

[59] Häring, Ivo, et al., *Towards a Generic Resilience Management, Quantification and Development Process: General Definitions, Requirements, Methods, Techniques and Measures, and Case Studies.* Association for Computing Machinery, 2017.

[60] A. A. Ganin, E. Massaro, A. Gutfraind, N. Steen, J. M. Keisler, A. Kott, R. Mangoubi, and I. Linkov, "Operational resilience: concepts, design and analysis," *Scientific Reports*, vol. 6, p. 19540, 2016.

[61] G. P. Cimellaro, A. M. Reinhorn, and M. Bruneau, "Framework for analytical quantification of disaster resilience," *Engineering Structures*, vol. 32, no. 11, pp. 3639–3649, 2010.

[62] D. Henry and J. Emmanuel Ramirez-Marquez, "Generic metrics and quantitative approaches for system resilience as a function of time," *Reliability Engineering & System Safety*, vol. 99, pp. 114–122, 2012.

[63] B. Deborah J., G. Richard D., M. Rosalie M., and W. John, "Cyber resiliency metrics, measures of effectiveness, and scoring," 2018. [Online]. Available: https://www.mitre.org/sites/default/files/publications/pr-18-2579-cyber-resiliency-metrics-measures-of-effectiveness-and-scoring.pdf

[64] E. Hollnagel, J. Pariès, D. Woods, and J. Wreathall, Eds., *Resilience Engineering in Practice: A Guidebook.* CRC Press, 2011, vol. 99.

[65] E. Bellini, L. Coconea, and P. Nesi, "A functional resonance analysis method driven resilience quantification for socio-technical systems," *IEEE Systems Journal*, 2019.

[66] E. Bellini, P. Bellini, D. Cenni, P. Nesi, G. Pantaleo, I. Paoli, and M. Paolucci, "An ioe and big multimedia data approach for urban transport system resilience management in smart cities," *Sensors*, vol. 21, no. 2, 2021.

[67] J. Andersson, V. Grassi, R. Mirandola, and D. Perez-Palacin, "A conceptual framework for resilience: fundamental definitions, strategies and metric," *Computing*, vol. 4, no. 103, pp. 559–588, 2021.

[68] ENISA, "Measurement frameworks and metrics for resilient networks and services - technical report," 2011. [Online]. Available: https://www.enisa.europa.eu/publications/metrics-tech-report/@@download/fullReport

[69] ENISA, "Artificial Intelligence Cybersecurity Challenges," 2020. [Online]. Available: https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges

[70] S. Zeadally, E. Adi, Z. Baig, and I. A. Khan, "Harnessing artificial intelligence capabilities to improve cybersecurity," *IEEE Access*, vol. 8, pp. 23 817–23 837, 2020.

[71] Z. Zhang, H. Ning, F. Shi, F. Farha, Y. Xu, J. Xu, F. Zhang, and K.-K. R. Choo, "Artificial intelligence in cyber security: research advances, challenges, and opportunities," *Artificial Intelligence Review*, pp. 1–25, 2021.

[72] J. E. Rubio, C. Alcaraz, R. Roman, and J. Lopez, "Current cyber-defense trends in industrial control systems," *Computers & Security Journal*, vol. 87, 11/2019 2019.

[73] T. Zoppi, A. Ceccarelli, and A. Bondavalli, "MADneSs: A Multi-Layer Anomaly Detection Framework for Complex Dynamic Systems," *IEEE Transactions on Dependable and Secure Computing*, vol. 18, no. 2, pp. 796–809, 2021.

[74] A. Aldweesh, A. Derhab, and A. Z. Emam, "Deep learning approaches for anomaly-based intrusion detection systems: A survey, taxonomy, and open issues," *Knowledge-Based Systems*, vol. 189, p. 105124, 2020.

[75] W. Li, W. Meng, and L. F. Kwok, "Surveying trust-based collaborative intrusion detection: State-of-the-art, challenges and future directions," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 1, pp. 280–305, 2021.

[76] E. Vasilomanolakis, S. Karuppayah, M. Muhlhauser, and M. Fischer, "Taxonomy and survey of collaborative intrusion detection," *ACM Computing Surveys*, vol. 47, no. 4, pp. 1–33, 2015.

[77] E. M. Campos, P. F. Saura, A. González-Vidal, J. L. Hernández-Ramos, J. B. Bernabe, G. Baldini, and A. Skarmeta, "Evaluating federated learning for intrusion detection in internet of things: Review and challenges," *Computer Networks*, p. 108661, 2021.

[78] J. Cumplido, C. Alcaraz, and J. Lopez, "Collaborative anomaly detection system for charging stations," in *Computer Security – ESORICS 2022*, V. Atluri, R. Di Pietro, C. D. Jensen, and W. Meng, Eds. Cham: Springer Nature Switzerland, 2022, pp. 716–736.

[79] U. Ünal, C. N. Kahya, Y. Kurtlutepe, and H. Dağ, "Investigation of cyber situation awareness via siem tools: a constructive review," in *2021 6th International Conference on Computer Science and Engineering (UBMK)*. IEEE, 2021, pp. 676–681.

[80] M. Conti, T. Dargahi, and A. Dehghantanha, "Cyber threat intelligence: challenges and opportunities," in *Cyber Threat Intelligence*. Springer, 2018, pp. 1–6.

[81] J. E. Rubio, C. Alcaraz, R. Rios, R. Roman, and J. Lopez, "Distributed detection of apts: Consensus vs. clustering," in *25th European Symposium on Research in Computer Security (ESORICS 2020)*, vol. 12308, 09/2020 2020, pp. 174–192.

[82] L. Cazorla, C. Alcaraz, and J. Lopez, "Awareness and reaction strategies for critical infrastructure protection," *Computers and Electrical Engineering*, vol. 47, pp. 299–317, 2015.

[83] O. Veledar, V. Damjanovic-Behrendt, and G. Macher, "Digital twins for dependability improvement of autonomous driving," in *Systems, Software and Services Process Improvement*, A. Walker, R. V. O'Connor, and R. Messnarz, Eds. Cham: Springer International Publishing, 2019, pp. 415–426.

[84] S. Almeaibed, S. Al-Rubaye, A. Tsourdos, and N. P. Avdelidis, "Digital twin analysis to promote safety and security in autonomous vehicles," *IEEE Communications Standards Magazine*, vol. 5, no. 1, pp. 40–46, 2021.

[85] J. Douthwaite, B. Lesage, M. Gleirscher, R. Calinescu, J. M. Aitken, R. Alexander, and J. Law, "A modular digital twinning framework for safety assurance of collaborative robotics," *Frontiers in Robotics and AI*, vol. 8, 2021. [Online]. Available: https://www.frontiersin.org/article/10.3389/frobt.2021.758099

[86] X. Zheng, J. Lu, and D. Kiritsis, "The emergence of cognitive digital twin: vision, challenges and opportunities," *International Journal of Production Research*, vol. 0, no. 0, pp. 1–23, 2021.

[87] N. Rajabli, F. Flammini, R. Nardone, and V. Vittorini, "Software verification and validation of safe autonomous cars: A systematic literature review," *IEEE Access*, vol. 9, pp. 4797–4819, 2021.

[88] A. Salvi, P. Spagnoletti, and N. S. Noori, "Cyber-resilience of critical cyber infrastructures: Integrating digital twins in the electric power ecosystem," *Computers & Security*, vol. 112, p. 102507, 2022.

[89] E. Bellini, F. Bagnoli, M. Caporuscio, E. Damiani, F. Flammini, I. Linkov, P. Liò, and S. Marrone, "Resilience learning through self adaptation in digital twins of human-cyber-physical systems," in *2021 IEEE International Conference on Cyber Security and Resilience (CSR)*, 2021, pp. 168–173.

[90] E. Brucherseifer, H. Winter, A. Mentges, M. Mühlhäuser, and M. Hellmann, "Digital Twin conceptual framework for improving critical infrastructure resilience," *at - Automatisierungstechnik*, vol. 69, no. 12, pp. 1062–1080, 2021. [Online]. Available: https://doi.org/10.1515/auto-2021-0104

[91] D. Holmes, M. Papathanasaki, L. Maglaras, M. A. Ferrag, S. Nepal, and H. Janicke, "Digital Twins and Cyber Security – solution or challenge?" in *2021 6th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM)*, 2021, pp. 1–8.

[92] R. Faleiro, L. Pan, S. Pokhrel, and R. Doss, "Digital Twin for Cybersecurity: Towards Enhancing Cyber Resilience," in *Broadband Communications, Networks, and Systems*, ser. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, W. Xiang, F. Han, and T. Phan, Eds., vol. 413. pringer, Cham, 2022.

[93] C. Alcaraz and J. Lopez, "Digital twin: A comprehensive survey of security threats," *IEEE Communications Surveys & Tutorials*, vol. 24, no. thirdquarter 2022, pp. 1475 – 1503, 04/2022 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9765576

[94] M. Caporuscio, F. Flammini, N. Khakpour, P. Singh, and J. Thornadtsson, "Smart-troubleshooting connected devices: Concept, challenges and opportunities," *Future Generation Computer Systems*, vol. 111, pp. 681–697, 2020.

[95] R. Frei, R. McWilliam, B. Derrick, A. Purvis, A. Tiwari, and G. Di Marzo Serugendo, "Self-healing and self-repairing technologies," *International Journal of Advanced Manufacturing Technology*, vol. 69, no. 5-8, pp. 1033–1061, 2013.

[96] L. Gazzola, D. Micucci, and L. Mariani, "Automatic software repair: A survey," *IEEE Transactions on Software Engineering*, vol. 45, no. 1, pp. 34–67, 2019.

[97] M. Huebscher and J. McCann, "A survey of autonomic computing - degrees, models, and applications," *ACM Computing Surveys*, vol. 40, no. 3, 2008.

[98] J. Kephart and D. Chess, "The vision of autonomic computing," *IEEE Computer*, vol. 36, no. 1, pp. 41–50, 2003.

[99] P. Singh, M. Saman Azari, F. Vitale, F. Flammini, N. Mazzocca, M. Caporuscio, and J. Thornadtsson, "Using log analytics and process mining to enable self-healing in the internet of things," *Environment Systems and Decisions*, vol. 42, no. 2, p. 234–250, 2022.

[100] N. Gol Mohammadi, *Trustworthiness-by-design*. Wiesbaden: Springer, 2019, pp. 79–118.