

IEEE Standard for Transparency of Autonomous Systems

IEEE Vehicular Technology Society

IEEE Robotics and Automation Society

Developed by the
Intelligent Transportation Systems Committee
and the
Standing Committee for Standards

IEEE Std 7001™-2021

STANDARDS

IEEE Standard for Transparency of Autonomous Systems

Developed by the

Intelligent Transportation Systems Committee
of the
IEEE Vehicular Technology Society

and the

Standing Committee for Standards
of the
IEEE Robotics and Automation Society

Approved 8 December 2021

IEEE SA Standards Board

Abstract: Measurable, testable levels of transparency, so that autonomous systems can be objectively assessed, and levels of compliance determined, are described in this standard.

Keywords: autonomous systems, artificial intelligence, ethics, IEEE 7001™, transparency

The Institute of Electrical and Electronics Engineers, Inc.
3 Park Avenue, New York, NY 10016-5997, USA

Copyright © 2022 by The Institute of Electrical and Electronics Engineers, Inc.
All rights reserved. Published 4 March 2022. Printed in the United States of America.

IEEE is a registered trademark in the U.S. Patent & Trademark Office, owned by The Institute of Electrical and Electronics Engineers, Incorporated.

PDF: ISBN 978-1-5044-8311-7 STD25178
Print: ISBN 978-1-5044-8312-4 STDPD25178

IEEE prohibits discrimination, harassment, and bullying.

For more information, visit <https://www.ieee.org/about/corporate/governance/p9-26.html>.

No part of this publication may be reproduced in any form, in an electronic retrieval system or otherwise, without the prior written permission of the publisher.

Important Notices and Disclaimers Concerning IEEE Standards Documents

IEEE Standards documents are made available for use subject to important notices and legal disclaimers. These notices and disclaimers, or a reference to this page (<https://standards.ieee.org/ipr/disclaimers.html>), appear in all standards and may be found under the heading “Important Notices and Disclaimers Concerning IEEE Standards Documents.”

Notice and Disclaimer of Liability Concerning the Use of IEEE Standards Documents

IEEE Standards documents are developed within the IEEE Societies and the Standards Coordinating Committees of the IEEE Standards Association (IEEE SA) Standards Board. IEEE develops its standards through an accredited consensus development process, which brings together volunteers representing varied viewpoints and interests to achieve the final product. IEEE Standards are documents developed by volunteers with scientific, academic, and industry-based expertise in technical working groups. Volunteers are not necessarily members of IEEE or IEEE SA, and participate without compensation from IEEE. While IEEE administers the process and establishes rules to promote fairness in the consensus development process, IEEE does not independently evaluate, test, or verify the accuracy of any of the information or the soundness of any judgments contained in its standards.

IEEE makes no warranties or representations concerning its standards, and expressly disclaims all warranties, express or implied, concerning this standard, including but not limited to the warranties of merchantability, fitness for a particular purpose and non-infringement. In addition, IEEE does not warrant or represent that the use of the material contained in its standards is free from patent infringement. IEEE standards documents are supplied “AS IS” and “WITH ALL FAULTS.”

Use of an IEEE standard is wholly voluntary. The existence of an IEEE Standard does not imply that there are no other ways to produce, test, measure, purchase, market, or provide other goods and services related to the scope of the IEEE standard. Furthermore, the viewpoint expressed at the time a standard is approved and issued is subject to change brought about through developments in the state of the art and comments received from users of the standard.

In publishing and making its standards available, IEEE is not suggesting or rendering professional or other services for, or on behalf of, any person or entity, nor is IEEE undertaking to perform any duty owed by any other person or entity to another. Any person utilizing any IEEE Standards document, should rely upon his or her own independent judgment in the exercise of reasonable care in any given circumstances or, as appropriate, seek the advice of a competent professional in determining the appropriateness of a given IEEE standard.

IN NO EVENT SHALL IEEE BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO: THE NEED TO PROCURE SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE PUBLICATION, USE OF, OR RELIANCE UPON ANY STANDARD, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE AND REGARDLESS OF WHETHER SUCH DAMAGE WAS FORESEEABLE.

Translations

The IEEE consensus development process involves the review of documents in English only. In the event that an IEEE standard is translated, only the English version published by IEEE is the approved IEEE standard.

Official statements

A statement, written or oral, that is not processed in accordance with the IEEE SA Standards Board Operations Manual shall not be considered or inferred to be the official position of IEEE or any of its committees and shall not be considered to be, nor be relied upon as, a formal position of IEEE. At lectures, symposia, seminars, or educational courses, an individual presenting information on IEEE standards shall make it clear that the presenter's views should be considered the personal views of that individual rather than the formal position of IEEE, IEEE SA, the Standards Committee, or the Working Group.

Comments on standards

Comments for revision of IEEE Standards documents are welcome from any interested party, regardless of membership affiliation with IEEE or IEEE SA. However, **IEEE does not provide interpretations, consulting information, or advice pertaining to IEEE Standards documents.**

Suggestions for changes in documents should be in the form of a proposed change of text, together with appropriate supporting comments. Since IEEE standards represent a consensus of concerned interests, it is important that any responses to comments and questions also receive the concurrence of a balance of interests. For this reason, IEEE and the members of its Societies and Standards Coordinating Committees are not able to provide an instant response to comments, or questions except in those cases where the matter has previously been addressed. For the same reason, IEEE does not respond to interpretation requests. Any person who would like to participate in evaluating comments or in revisions to an IEEE standard is welcome to join the relevant IEEE working group. You can indicate interest in a working group using the Interests tab in the Manage Profile and Interests area of the [IEEE SA myProject system](#). An IEEE Account is needed to access the application.

Comments on standards should be submitted using the [Contact Us](#) form.

Laws and regulations

Users of IEEE Standards documents should consult all applicable laws and regulations. Compliance with the provisions of any IEEE Standards document does not constitute compliance to any applicable regulatory requirements. Implementers of the standard are responsible for observing or referring to the applicable regulatory requirements. IEEE does not, by the publication of its standards, intend to urge action that is not in compliance with applicable laws, and these documents may not be construed as doing so.

Data privacy

Users of IEEE Standards documents should evaluate the standards for considerations of data privacy and data ownership in the context of assessing and using the standards in compliance with applicable laws and regulations.

Copyrights

IEEE draft and approved standards are copyrighted by IEEE under US and international copyright laws. They are made available by IEEE and are adopted for a wide variety of both public and private uses. These include both use, by reference, in laws and regulations, and use in private self-regulation, standardization, and the promotion of engineering practices and methods. By making these documents available for use and adoption by public authorities and private users, IEEE does not waive any rights in copyright to the documents.

Photocopies

Subject to payment of the appropriate licensing fees, IEEE will grant users a limited, non-exclusive license to photocopy portions of any individual standard for company or organizational internal use or individual, non-commercial use only. To arrange for payment of licensing fees, please contact Copyright Clearance Center, Customer Service, 222 Rosewood Drive, Danvers, MA 01923 USA; +1 978 750 8400; <https://www.copyright.com/>. Permission to photocopy portions of any individual standard for educational classroom use can also be obtained through the Copyright Clearance Center.

Updating of IEEE Standards documents

Users of IEEE Standards documents should be aware that these documents may be superseded at any time by the issuance of new editions or may be amended from time to time through the issuance of amendments, corrigenda, or errata. An official IEEE document at any point in time consists of the current edition of the document together with any amendments, corrigenda, or errata then in effect.

Every IEEE standard is subjected to review at least every 10 years. When a document is more than 10 years old and has not undergone a revision process, it is reasonable to conclude that its contents, although still of some value, do not wholly reflect the present state of the art. Users are cautioned to check to determine that they have the latest edition of any IEEE standard.

In order to determine whether a given document is the current edition and whether it has been amended through the issuance of amendments, corrigenda, or errata, visit [IEEE Xplore](#) or [contact IEEE](#). For more information about the IEEE SA or IEEE's standards development process, visit the IEEE SA Website.

Errata

Errata, if any, for all IEEE standards can be accessed on the [IEEE SA Website](#). Search for standard number and year of approval to access the web page of the published standard. Errata links are located under the Additional Resources Details section. Errata are also available in [IEEE Xplore](#). Users are encouraged to periodically check for errata.

Patents

IEEE Standards are developed in compliance with the [IEEE SA Patent Policy](#).

Attention is called to the possibility that implementation of this standard may require use of subject matter covered by patent rights. By publication of this standard, no position is taken by the IEEE with respect to the existence or validity of any patent rights in connection therewith. If a patent holder or patent applicant has filed a statement of assurance via an Accepted Letter of Assurance, then the statement is listed on the IEEE SA Website at <https://standards.ieee.org/about/sasb/patcom/patents.html>. Letters of Assurance may indicate whether the Submitter is willing or unwilling to grant licenses under patent rights without compensation or under reasonable rates, with reasonable terms and conditions that are demonstrably free of any unfair discrimination to applicants desiring to obtain such licenses.

Essential Patent Claims may exist for which a Letter of Assurance has not been received. The IEEE is not responsible for identifying Essential Patent Claims for which a license may be required, for conducting inquiries into the legal validity or scope of Patents Claims, or determining whether any licensing terms or conditions provided in connection with submission of a Letter of Assurance, if any, or in any licensing agreements are reasonable or non-discriminatory. Users of this standard are expressly advised that determination of the validity of any patent rights, and the risk of infringement of such rights, is entirely their own responsibility. Further information may be obtained from the IEEE Standards Association.

IMPORTANT NOTICE

IEEE Standards do not guarantee or ensure safety, security, health, or environmental protection, or ensure against interference with or from other devices or networks. IEEE Standards development activities consider research and information presented to the standards development group in developing any safety recommendations. Other information about safety practices, changes in technology or technology implementation, or impact by peripheral systems also may be pertinent to safety considerations during implementation of the standard. Implementers and users of IEEE Standards documents are responsible for determining and complying with all appropriate safety, security, environmental, health, and interference protection practices and all applicable laws and regulations.

Participants

At the time this IEEE standard was completed, the Autonomous Systems Validation Working Group had the following membership:

Alan F.T. Winfield, *Chair*
Eleanor “Nell” Watson, *Vice Chair*
Takashi Egawa, *Secretary*

Emily Barwell	Naomi Jacobs	Fahime Rajabiyazdi
Iain Barclay	Milan Markovic	Randy K. Rannow
Serena Booth	Roderick I. Muttram	Andreas Theodorou
Louise A. Dennis	Lawrence Nadel	Mark A. Underwood
Helen Hastie	Iman Naja	Oskar von Stryk
Ali Hossaini	Joanna Olszewska	Robert H. Wortham

The following members of the individual Standards Association balloting group voted on this standard. Balloters may have voted for approval, disapproval, or abstention.

Robert Aiello	Edmund Kienast	Patty Polpattana
M. Victoria Alonso	Ansgar Koene	Venkatesha Prasad
Lyria Bennett Moses	Thomas Kurihara	Fahime Rajabiyazdi
Pieter Botman	Sean Laroque-Doherty	Randy K. Rannow
Bill Brown	Julio Leite	Annette Reilly
William Byrd	James Lepp	Robert Schaaf
Diego Chiozzi	Gerri Light	John Sheppard
Takashi Egawa	Juan Antonio Lloret	Gerald Stueve
Avraham Freedman	Egea	Andreas Theodorou
Paulo Goncalves	Lars Luenenburger	John Vergis
Louis Gullo	Javier Luiso	Ionel Marius Vladan
Didem Gurdur Broo	Milan Markovic	Oskar von Stryk
Marco Hernandez	Rajesh Murthy	Lei Wang
Ali Hessami	Roderick I. Muttram	Eleanor “Nell” Watson
Werner Hoelzl	Iman Naja	Alan F.T. Winfield
Dennis Holstein	Joanna Olszewska	Robert H. Wortham
Masao Ito	Satoshi Oyama	Hasan Yasar
Naomi Jacobs	Sivaraman P.	Naritoshi Yoshinaga
Piotr Karocki	Davy Pisssoort	Yu Yuan

When the IEEE SA Standards Board approved this standard on 8 December 2021, it had the following membership:

Gary Hoffman, *Chair*
Jon Walter Rosdahl, *Vice Chair*
John D. Kulick, *Past Chair*
Konstantinos Karachalios, *Secretary*

Edward A. Addy
Doug Edwards
Ramy Ahmed Fathy
J. Travis Griffith
Thomas Koshy
Joseph L. Koepfinger*
David J. Law

Howard Li
Daozhuang Lin
Kevin Lu
Daleep C. Mohla
Chenhui Niu
Damir Novosel
Annette Reilly
Dorothy Stanley

Mehmet Ulema
Lei Wang
F. Keith Waters
Karl Weber
Sha Wei
Howard Wolfman
Daidi Zhong

*Member Emeritus

Introduction

This introduction is not part of IEEE Std 7001™-2021, IEEE Standard for Transparency of Autonomous Systems.

IEEE Std 7001-2021, IEEE Standard on Transparency of Autonomous Systems, sets out measurable, testable levels of transparency for autonomous systems. The standard was inaugurated to help make actionable the principle that it should always be possible to understand why and how an autonomous system made a particular decision and the consequential system's behaviors. Transparency is one of the eight general principles set out in *IEEE Ethically Aligned Design* [B21],¹ stated as “The basis of a particular autonomous and intelligent system decision should always be discoverable.” A working group tasked with drafting this standard was proposed in direct response to a recommendation in the general principles section of *IEEE Ethically Aligned Design*.

The IEEE Project Authorization Request (PAR) was approved on 7 December 2016. The sponsor committees are VT/ITS—Intelligent Transportation Systems and the RAS/SC Standing Committee for Standards.

The IEEE 7000 series of IEEE standards have been developed in parallel with IEEE's ethics certification program for autonomous and intelligent systems and have benefitted from the pool of global expertise of IEEE.

The specific aim of the ethics artificial intelligence system (AIS) certification program has been to develop assessment criteria that assist duty holders with “self” or “independent” ethical scrutiny and assurance of products, services, and systems. The ethics AIS certification program's objectives are therefore complementary to the guidelines and requirements of the IEEE 7000 series of standardization projects and standards. In particular, the ethics AIS certification criteria are focused on the manifest and verifiable emergent properties/outcomes, whereas our standards generally prescribe processes for the realization of a range of ethical attributes.

The IEEE certification program on ethics AIS and the IEEE 7000 series of technology ethics standards provide a comprehensive best practice and voluntary toolkit for responsible ethically aligned design and deployment of autonomous and intelligent systems.

For more information visit: <https://ethicsinaction.ieee.org/p7000/>.

¹The numbers in brackets correspond to those of the bibliography in Annex C.

Contents

1. Overview	11
1.1 Scope	11
1.2 Purpose	12
1.3 Target audience	12
1.4 Approaches to transparency	13
1.5 How to apply this standard	13
1.6 Word usage	14
2. Normative references	14
3. Definitions, acronyms, and abbreviations	14
3.1 Definitions	14
3.2 Acronyms and abbreviations	15
4. Key concepts	15
4.1 System transparency and explainability	15
4.2 System autonomy	16
5. Transparency requirements by stakeholder and level	18
5.1 Stakeholders who benefit directly from increased transparency	18
5.2 Expert stakeholders who require transparency as part of their work	22
Annex A (informative) A guide on how and when to use this standard	31
Annex B (informative) Scenarios	36
Annex C (informative) Bibliography	51

IEEE Standard for Transparency of Autonomous Systems

1. Overview

1.1 Scope

This standard is broadly applicable to all autonomous systems, including both physical and non-physical systems. Examples of the former include vehicles with automated driving systems or assisted living (care) robots. Examples of the latter include medical diagnosis (recommender) systems or chatbots. Of particular interest to this standard are autonomous systems that have the potential to cause harm. Safety-critical systems are therefore within scope. This standard considers systems that have the capacity to directly cause either physical, psychological, societal, economic or environmental, or reputational harm, as within scope. Harm might also be indirect, such as unauthorized persons gaining access to confidential data or “victimless crimes” that affect no-one in particular yet have an impact upon society or the environment.

Intelligent autonomous systems that use machine learning are also within scope. The data sets used to train such systems are also within the scope of this standard when considering the transparency of the system as a whole.

This standard provides a framework to help developers of autonomous systems both review and, if needed, design features into those systems to make them more transparent. The framework sets out requirements for those features, the transparency they bring to a system, and how they would be demonstrated in order to determine conformance with this standard.

Future standards may choose to focus on specific applications or technology domains. This standard is intended as an “umbrella” standard from which domain-specific standards might develop (for instance, standards for transparency in autonomous vehicles, medical or healthcare technologies, etc.).

This standard does not provide the designer with advice on how to design transparency into their system. Instead, it defines a set of testable levels of transparency and a standard set of requirements that shall be met in order to satisfy each of these levels.

Transparency cannot be assumed. An otherwise well-designed system may not be transparent. Many well-designed systems are not transparent. Autonomous systems, and the processes by which they are designed, validated, and operated, will only be transparent if this is designed into them. In addition, methods for testing, measuring, and comparing different levels of transparency in different systems are needed.

Note that system-system transparency (transparency of one system to another) is out of scope for this standard. However, this document does address the transparency of the engineering process. Transparency regarding how subsystems within an autonomous system interact is also within the scope of this standard.

1.2 Purpose

The purpose of this standard is to set out measurable, testable levels of transparency for autonomous systems. The general principle behind this standard is that it should always be possible to understand why and how the system behaved the way it did. Transparency is one of the eight General Principles set out in *IEEE Ethically Aligned Design* [B21], stated as “The basis of a particular autonomous and intelligent system decision should always be discoverable.” A working group tasked with drafting this standard was set up in direct response to a recommendation in the general principles section of *IEEE Ethically Aligned Design*.

There are several reasons transparency is important:

- Modern autonomous systems are designed to work with or alongside humans who need to be able to understand what the systems are doing and why. Imagine a care robot that behaves in a way that is puzzling or unpredictable. Persons that interact with the robot and their wardens may be less likely to have confidence in the robot, therefore they will be less likely to make full use of it. Transparency is important in adjusting expectations and, hence, building confidence.
- Autonomous systems can sometimes fail. If physical robots fail, they can cause physical harm or injury. Failure of non-physical (software) systems can also cause harm. A medical diagnosis artificial intelligence system (AIS) might, for instance, give the wrong diagnosis, or a credit scoring AIS might make an incorrect recommendation and cause a person’s loan application to be rejected. Without transparency, finding out what went wrong and why is extremely difficult and may, in some cases, be impossible. Equally, finding out how and why a system made a correct decision is important for the processes of verification and validation.
- Without transparency, accountability and the attribution of responsibility can be difficult. Public confidence in technology requires both transparency and accountability. Transparency is needed so that the public can understand who is responsible for the way autonomous systems work and—equally importantly—sometimes do not work. It might also be important to establish who is responsible for insurance or regulatory purposes or in an administrative proceeding or court of law. Transparency improves accountability, which might in turn support judicial processes. Finally, following high profile accidents, society can benefit from the reassurance of knowing that problems have been found and addressed.

1.3 Target audience

The target audience of this standard are those designers, developers, builders, maintainers, and operators, as well as decision-makers and procurers in organizations using and deploying autonomous systems (collectively, “designers”) of autonomous systems who either wish to or are required to engineer systems that have a certain degree of transparency. This standard can help designers to self-assess the transparency of their system and then provide recommendations for additional transparency measures if necessary. The standard can also help transparency requirements to be specified in such a way that conformance can be demonstrated.

A secondary audience for this standard are groups who benefit from transparency. These groups are referred to as stakeholders. There are two groups of stakeholders:

- Stakeholders who benefit directly from increased transparency—these include both direct users of autonomous systems and wider society (see 5.1).
- Expert stakeholders who require transparency as part of their work—these include certification or regulatory bodies, incident/accident investigators, and expert advisors in administrative actions or litigation (see 5.2).

1.4 Approaches to transparency

Broadly, transparency requires three parallel approaches, as follows:

- The first is process standards for ethically aligned design; that is, standards setting out human processes for ethically designing, validating, and operating robotics and AI systems. The IEEE Standards Association working groups are currently drafting a series of so-called human standards. The first of these, IEEE Std 7000™-2021, [B25], is a model process for addressing ethical concerns during system design.
- Second, a standard is needed for transparency; IEEE Std 7001-2021 is that standard.
- Third, technologies for transparency are needed. This standard does not specify technologies to support transparency, although for one stakeholder group, incident/accident investigators, this standard requires data logging to be incorporated into autonomous systems. Data logging is required to provide investigators with time stamped records of what a system was doing prior to and during an incident. The technical specification of such data logging systems is outside the scope of this standard.

Transparency has widespread economic and social benefits, such as greater social trust. Greater transparency eases coordination through sharing of information such as plans, intentions, and status. Transparency can inform consumer choice, thereby rewarding quality and excellence, and encourages less scrupulous actors to change their behavior. Transparency also allows incentives to be aligned more easily. For example, insurers may be able to offer a more accurate premium if they better understand the characteristics of an autonomous system in its operation and not merely after an incident.

However, transparency should be designed into the system; ideally from its inception rather than retroactively. The quality of transparency does not manifest without careful consideration and adherence to best practices and rigorous standards.

1.5 How to apply this standard

There are two ways in which this standard can be applied in practice, as follows:

- A System Transparency Assessment (STA) is the process of evaluating the transparency of an existing autonomous system, for each stakeholder group.

A system is conformant with IEEE Std 7001-2021 if the STA determines that it meets at least Transparency Level 1 in at least one declared stakeholder group. Such minimal conformance may not be acceptable to the stakeholders of the system in question. Determination of what are appropriate or minimum acceptable levels of transparency for a given system is made by writing a System Transparency Specification (STS), as defined in the next list item. Direct comparison of transparency requirements in the STS with measured transparency in the STA can help reveal transparency gaps that need to be addressed. Information that improves transparency shall also be provided in an accessible format that supports comprehension by stakeholders.
- An STS is the process of defining the transparency requirements of an autonomous system, for each stakeholder group. An STS may be written at any time during a system's lifecycle, though the best and expected practice would be to specify transparency requirements prior to system design (see IEEE/ISO/IEC Std 15288:2015 [B26] and IEEE/ISO/IEC Std 12207:2017 [B30]).

It is important to note that transparency requirements will vary considerably from one system to another. A prerequisite of writing an STS is to decide on the appropriate level of transparency for each stakeholder group and for the system under consideration.

Detailed guidelines on how and when to apply this standard, with templates for the processes of STA and STS, are given in [Annex A](#). This standard does not prescribe minimum acceptable levels of transparency for particular autonomous systems (or categories of systems), however, detailed worked examples of STA and STS are given in a set of scenarios, for both fictional and (some) real autonomous systems, in [Annex B](#).

1.6 Word usage

The word *shall* indicates mandatory requirements strictly to be followed in order to conform to the standard and from which no deviation is permitted (*shall* equals *is required to*).^{2,3}

The word *should* indicates that among several possibilities one is recommended as particularly suitable, without mentioning or excluding others; or that a certain course of action is preferred but not necessarily required (*should* equals *is recommended that*).

The word *may* is used to indicate a course of action permissible within the limits of the standard (*may* equals *is permitted to*).

The word *can* is used for statements of possibility and capability, whether material, physical, or causal (*can* equals *is able to*).

2. Normative references

The following referenced documents are indispensable for the application of this document (i.e., they must be understood and used, so each referenced document is cited in text and its relationship to this document is explained). For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments or corrigenda) applies.

IEC/IEEE 82079-1, International Standard for Preparation of information for use (instructions for use) of products—Part 1: Principles and general requirements.^{4,5,6}

3. Definitions, acronyms, and abbreviations

3.1 Definitions

For the purposes of this document, the following terms and definitions apply. The *IEEE Standards Dictionary Online* should be consulted for terms not defined in this clause.⁷

autonomous system: A system that has the capacity to make decisions itself in response to some input data or stimulus with a varying degree of human oversight or intervention depending on the system's level of autonomy.

domain expert users: Persons who carry some responsibility for how an autonomous system is used or are responsible for operating and supervising autonomous systems.

²The use of the word *must* is deprecated and cannot be used when stating mandatory requirements, *must* is used only to describe unavoidable situations.

³The use of *will* is deprecated and cannot be used when stating mandatory requirements, *will* is only used in statements of fact.

⁴IEC publications are available from the International Electrotechnical Commission (<https://www.iec.ch/>). IEC publications are also available in the United States from the American National Standards Institute (<http://www.ansi.org>).

⁵The IEEE standards or products referred to in this clause are trademarks of The Institute of Electrical and Electronics Engineers, Inc.

⁶IEEE publications are available from The Institute of Electrical and Electronics Engineers, 445 Hoes Lane, Piscataway, NJ 08854, USA (<https://standards.ieee.org/>).

⁷*IEEE Standards Dictionary Online* is available at: <http://dictionary.ieee.org>. An IEEE Account is required for access to the dictionary, and one can be created at no charge on the dictionary sign-in page.

explainability: The extent to which the information made transparently available to a stakeholder can be readily interpreted and understood by a stakeholder.

non-expert users: Persons who have only a brief interaction or who interact every day with an autonomous system.

stakeholders: An individual or organization having a right, share, claim, or interest in a system or in its possession of characteristics that meet their needs and expectations.

superusers: Experts not only in autonomous systems but also in the particular systems for which they are responsible. *See also:* **domain expert users.**

System Transparency Assessment (STA): The process of evaluating the transparency of an existing autonomous system, for each stakeholder group.

System Transparency Specification (STS): The process of defining the transparency requirements of an autonomous system for each stakeholder group.

transparency: A transfer of information from an autonomous system or its designers to a stakeholder that is truthful; contains information relevant to the causes of some action, decision, or behavior; and is presented at a level of abstraction and in a form meaningful to the stakeholder. Transparency should be mindful of the stakeholders' likely perception and comprehension, and should avoid disclosing information in a manner that, while technically true, is framed in a way that leads to misapprehension.

3.2 Acronyms and abbreviations

AIS	artificial intelligence system
GDPR	General Data Protection Regulation
Med DSS	medical decision support system
NLP	natural language processing
STA	System Transparency Assessment
STS	System Transparency Specification

4. Key concepts

4.1 System transparency and explainability

The principle behind this standard is that it should always be possible to understand why and how (e.g., by what decision-making logic, algorithm, or prediction mechanism) an autonomous system behaved in a particular way.

In this document, the term transparency refers to a transfer of information from an autonomous system, or its configurers, operators, designers and developers to a stakeholder. Such information shall be truthful; contain information relevant to the causes of some action, decision, or behavior; and be presented at a level of abstraction and in a form (typically natural language) meaningful to the stakeholder. Such information can be offered both to account for past behavior and to describe potential future behavior, as well as to expose the capabilities and limitations of a system.

To consider an autonomous system transparent for inspection, the stakeholder should have the ability to request meaningful explanations of the system's status, either at a specific moment or over a specific period or of the general principles by which decisions are made (as appropriate to the stakeholder) (see Theodorou, Wortham, and Bryson, [B56]).

The system's status shall include relevant goals; progress in relation to those goals; models of its past, current, and potential future environmental context (from sensors and other information); and relevant information about its current performance, such as reliability and error messages (see Wortham, Theodorou, and Bryson [B63]). For an autonomous system to be considered transparent, this information shall be presented in a human understandable form.

However, a developer may not be able or may not wish to achieve the same degree of transparency in all systems; for instance, non-expert users likely do not need logs of sensor inputs whereas incident investigators are likely to need precisely such information. Transparency is a quality that enables technical experts such as designers, testers, behavioral analysts and incident investigators to access data from a system that describes the process behind its decisions and behaviors.

Thus, this standard defines different levels of transparency based on the system itself and the stakeholder accessing the transparent information. Some of these levels (all levels for some stakeholders) require the system to be explainable, not just transparent, in order to conform with this standard.

A system that is explainable is said to have the quality of explainability. Explainability describes the extent to which the information made transparently available to a stakeholder can be readily interpreted by that stakeholder. Explainability is defined as the extent to which the internal state and decision-making processes of an autonomous system are accessible to non-expert stakeholders. Such explanations could be generated either by the system itself, or by a separate (machine) interpreter. Explainability requires being able to describe the causality behind a system's actions, at some level of abstraction appropriate to a non-expert.

It should be noted that the terms transparency and explainability are used in many subfields of artificial intelligence, robotics, and autonomous systems with slightly different meanings. Our intent here is to define their usage within this standard, not to mandate or prescribe their usage elsewhere. In particular, it is noted that in many areas of AI and robotics transparency refers to what this document refers to as *explainability*. In other words, it refers to the provision of information in a form readily understandable by a stakeholder and, indeed, the concept of explainability as defined here draws on these definitions. Similarly, it is noted that there are fields in which the term transparency implies that the system has become invisible to the user so that they feel they are directly controlling a task of the system (see Sheridan and Verplank [B51]).

Transparency is necessary but not sufficient for reducing the risk of psychological harm or distress. Explainability is a crucial additional factor for building trust and assurance between an autonomous system and its end-users or members of the public. It is also important to note that providing an explanation does not necessarily make a system's actions completely transparent (see De Graaf and Malle [B13]).

4.2 System autonomy

For the purpose of this standard, an autonomous system is defined as a system that has the capacity to make decisions itself in response to some input data or stimulus with a varying degree of human intervention, depending on the system's level of autonomy.

System autonomy falls on a spectrum from zero to full autonomy, where zero means the system is entirely under human control and full autonomy means the system can accomplish a goal without human guidance or intervention.

Levels of autonomy are included in this section in order to emphasize that "autonomous systems" addressed in this standard is a superset that includes semi-autonomous or supervised autonomous systems (which describe most extant systems).

There are many definitions in the literature for degrees (or levels) of autonomy. Sheridan [B50] defined 10 levels of autonomy from level 1, i.e., "computer offers no assistance," to level 10, i.e., "computer does everything even ignoring the human." Endsley and Kaber [B18] similarly defined 10 levels from level 1, i.e., "manual control," to level 10, i.e., "full automation in which the system carries out all actions and itself decides if it needs to suspend operation for human intervention."