IEEE SA
STANDARDS ASSOCIATION

IEEE Standard for Transparency of Autonomous Systems

STANDARDS

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

IEEE

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

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

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

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

Contents

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.

[2]The use of the word *must* is deprecated and cannot be used when stating mandatory requirements, *must* is used only to describe unavoidable situations.
[3]The use of *will* is deprecated and cannot be used when stating mandatory requirements, *will* is only used in statements of fact.
[4]IEC 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).
[5]The IEEE standards or products referred to in this clause are trademarks of The Institute of Electrical and Electronics Engineers, Inc.
[6]IEEE 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."

NIST introduced the Autonomy Levels for Unmanned Systems (ALFUS) as a nomenclature consisting of four levels of autonomy, namely, remote controlled, teleoperated, semi-autonomous, and fully autonomous (see NIST SP 1011-II-1.0 [B38]).

Based on ALFUS nomenclature, Durst and Gray [B16] expanded these four levels as follows:

a) *Human Operated:* A human operator makes all decisions.

b) *Human Delegated:* The system can perform many functions independently of human control when delegated to do so.

c) *Human Supervised:* The system can perform a wide variety of activities when given top-level permission or direction by a human.

d) *Fully Autonomous:* The system receives goals from humans and translates them into tasks to be performed without human interaction.

There are three components of supervised autonomy, as follows:

— Direction, i.e., telling a system what to do

— Monitoring, i.e., watching what the system is doing

— Control, i.e., being able to intervene and change what the system is doing

Regarding control, shared autonomy is a frequently used term to describe the situation where control of a machine is shared between a human operator and a computer system to achieve a goal, either remotely (as in Mercier and Tessier [B37]) or in the same shared space. In this situation, conflicts are likely to occur, and how easily these conflicts are resolved depends on the transparency of the machine's reasoning.

In IEEE Std 1872-2015, IEEE Standard on Ontologies for Robotics and Automation [B24], the definitions of the levels of autonomy follow the operation modes defined by the ALFUS nomenclature. Furthermore, IEEE Std 1872-2015 defines the automated attribute for systems acting as automata in a process, e.g., clockworks [B24].

For driverless cars, the Society of Automotive Engineers has defined six levels of autonomy from level 0, manually driven, to level 5, fully autonomous in all driving scenarios (SAE J3016_201806 [B45]).

It is worth noting the degree of autonomy of a system could vary depending on the scale of the system inspection. For example, a system could be semi-autonomous when completing an intended task, but could contain one or several autonomous sub-systems, e.g., relying on narrow artificial intelligence such as computer/machine vision processes, which can perform some sub-tasks autonomously (see Olszewska [B41]).

Furthermore, all of the systems this standard interact with humans in some way. For example, the system can make a recommendation to a human user on the basis of some digital input data, or in the case of a physical robot, make a decision about a course of action in response to sensor input data. Hence, in practice, no such system is 100% autonomous (i.e., self-determining), since all these systems are at some level commanded, monitored, and/or supervised by humans.

A further helpful reference in the context of Human-Robot Interaction is "Towards a framework for levels of robot autonomy in human-robot interaction," (Beer, Fisk, and Rogers [B4]). For a deeper and broader perspective, see also the MIT series *Intelligent Robotics and Autonomous Agents* [B3].

5. Transparency requirements by stakeholder and level

Requirements for measurable, testable levels of transparency are set within each stakeholder category. Levels of transparency are defined from 0 (no transparency) to 5 (the maximum achievable level of transparency). Each definition is a requirement, expressed as a qualitative property of the system that must be met. In each case, the test is simply that of determining whether the requirement is met or not, i.e., the transparency property required by a given level for a given stakeholder group is either demonstrably present or it is not. The choice of five levels is a compromise between a reasonable degree of granularity while allowing for discernible differences between successive levels.

Levels 1 to 5 have been defined to describe successively greater levels of transparency. All levels are judged to be technically feasible while each successive level is typically more challenging. For two categories of stakeholder, each level builds upon previous levels, so it is expected that when a system meets level *n* for a particular category, then it also meets levels *n* – 1, etc.

Stakeholder categories and their transparency definitions are independent of each other. There is no expectation that if a system meets level *n* in one category it will also meet the same level in other stakeholder categories. Levels that are not cumulative or categories that are not strictly independent are noted in 5.1 and 5.2. It should also be noted that any particular stakeholder may be interested in the transparency measures of other stakeholders for redundancy and cross-validation purposes.

Note that the levels of transparency set out in this clause are unrelated to the levels of autonomy in 4.2. Similarly, there is no expectation that higher-autonomy systems are required to conform with the higher levels of transparency in any of the categories below.

This clause is presented in two parts: Subclause 5.1 covers stakeholders who benefit directly from increased transparency and 5.2 covers expert stakeholders who require transparency as part of their work.

This standard recognizes but does not intend to restate or replace applicable laws and regulations regarding personal data, data privacy and data security. Users of this standard are responsible for referring to and observing all such laws and regulations. Conformance with the provisions of this standard does not imply conformance with any applicable legal or regulatory requirements.

5.1 Stakeholders who benefit directly from increased transparency

5.1.1 Users of autonomous systems

Autonomous systems shall provide a simple, understandable way for the user to understand what the system is doing and why and how the system is doing what it is doing. Not all users will require the same degree of system transparency; non-expert users will typically need simple and understandable high-level explanations of a system's decisions and actions, while expert users will require more complete and informative transparency.

The term user is defined as falling on a broad spectrum from non-expert users of autonomous systems to superusers, as follows:

— *Non-expert users* include both persons who have only a brief interaction with the system (for instance, when collecting a food delivery from an autonomous delivery robot or when using an automated hotel checking-in system) and persons who interact every day with the system (for instance, an assisted living robot, robot vacuum cleaner, or conversational AIS such as a smart speaker). Falling between non-expert users and superusers, is a category of domain expert users.

— *Domain expert users* include, for instance, a medical doctor using a medical diagnosis AIS as a diagnostic assistant in a clinical setting or a team of nuclear systems engineers supervising a semi-autonomous robot (or system of robots) to remotely repair or upgrade a reactor. Such domain expert

users carry some responsibility for how the system is used. The clinician, for instance, is responsible for interpreting the advice given by her diagnostic assistant. Similarly, the nuclear engineers are responsible for how the robots are deployed. This category also includes owner-drivers of autonomous vehicles as they too are responsible for the autonomous vehicle while its driver assist functions are engaged. Another group of domain expert users are those responsible for operating and supervising autonomous systems, for instance, those persons charged with managing and dispatching autonomous food delivery robots.

— *Superusers* are experts not only in autonomous systems but the particular systems for which they are responsible. Such superusers include persons responsible for development, fault diagnosis, repair, maintenance and upgrade, in addition to the operation and supervision, of particular autonomous systems.

It is noted that explaining current behavior/actions and explaining the system's general principles of operation are separate aspects of transparency. In defining transparency for users, it is necessary to be mindful of the importance of managing expectations of what the system can and cannot do in a way that does not confuse or upset the non-expert user.

For this category of stakeholder, the levels of transparency are not progressive, i.e., fulfillment of an earlier level is not necessary to achieve a higher one.

Transparency requirements for users are given in Table 1.

Table 1—Transparency requirements for users

Level	Definition
0 (lowest)	No transparency.
1	The user shall be provided with accessible[a] information that provides as a minimum the following: a) example scenarios with the expected and anticipated system behavior including degraded modes of operation and b) general principles of its operation, i.e., if there is a learning component and what data it uses. The documentation shall explain the system's general principles of operation. For a system that uses machine learning the documentation should provide a simple explanation of which sources the system examines/uses as part of the learning process, including any possible sources of bias. This documentation shall for example be in the form of a written manual, pictorial, or audio guide as appropriate to the user, which provides the user with an explanation of how the system behaves in the various circumstances and situations its designers expect it to encounter. Domain expert users and superusers shall be provided with user documentation as specified above and prepared in accordance with IEC/IEEE 82079-1[b]. This documentation shall detail the safe operation and supervision of the system. For superusers, the documentation shall additionally detail procedures for system fault diagnosis, repair, maintenance, upgrade, and end-of-life decommissioning.
2	The user shall be provided with interactive training material that allows the user to rehearse their interactions with the system in specific and relevant virtual situations. This interactive material shall be in the form of an interactive presentation, video, or simulation, which allows the user to rehearse their interactions with the system in specific different situations. In addition, domain expert users and superusers shall be provided with interactive training materials on the safe operation and supervision of the system. Superusers shall additionally be provided with interactive training materials covering fault diagnosis, repair, maintenance, upgrade, and end-of-life decommissioning.

Table continues

Table 1—Transparency requirements for users *(continued)*

Level	Definition
3	The non-expert user shall be provided with user-initiated functionality that produces a brief and immediate explanation of the system's most recent activity. These explanations shall be expressed through commonly understandable means such as natural language or another appropriate medium (e.g., a pictorial). Neither making requests nor understanding the system's responses to those requests shall require that the non-expert user undergo any training. However, advisories for safety or legal reasons are acceptable as may be necessary. *An example would be a robot or physical system equipped with a speech recognition system that will respond to the user asking, "Robot why did you just do that?" by producing—in plain language—a spoken explanation for its most recent action. For instance: "I stopped because I am programmed not to bump into you." An example of a non-physical system would be software in which either a touch screen button or a spoken request produces a similar explanation. An example of an advisory would be information on safety that must be understood prior to use, such as important safety information, or an age restriction.* For systems designed to be used by domain experts, the same functionality specified above shall be provided, except that a) the system shall allow explanations for any of its recent decisions to be requested and b) the explanations may be expressed using domain appropriate language. Domain experts shall additionally be provided with documentation detailing how these explanations should be requested and interpreted. Such documentation should also cover natural language processing (NLP) subsystems, if present. *An example would be a medical doctor using a medical diagnosis AIS as a diagnostic assistant. The system would allow the doctor to ask for an explanation of a recent recommendation, in language that allows the doctor to assess its plausibility.*
4	The non-expert user shall be provided with a user-initiated functionality that produces a brief and immediate explanation of what the system does in a given situation. Conformance with this level of transparency allows the user to explore hypothetical "what if" scenarios in a given situation, if applicable to the system's scope of work. Neither making requests nor understanding the system's responses to those requests shall require that the non-expert user undergo any training, though familiarization with the system's user documentation is required. *A robot or physical system should be able to respond to requests (possibly including gestures or eye contact) including both "Why did you just do that?" and "What would you do if .. xxx ..?" (for example "Robot what would you do if I fell down?" or "Robot what would you do if I forget to take my medicine?"), in natural language or equivalent signals.* *Non-physical systems should have an equivalent function, allowing the user to ask, "What would you decide/recommend if I asked you xxx, and why?"* For systems designed to be used by domain experts, the same functionality specified here shall be provided, except that the explanations may be expressed using domain appropriate language. Domain experts shall additionally be provided with documentation detailing how these explanations should be requested and interpreted. Such documentation should also cover NLP subsystems, if present. Importantly this level of transparency allows the user to explore counterfactuals (see Wachter, Mittelstadt, and Russell [B57]).
5 (highest)	The user shall be provided with a continuous explanation of behavior that adapts the content and presentation of the explanation based on the user's information needs and context. This shall include access to log files and training data as long as they do not contain sensitive information such as personal data. An explanation of operation shall be achieved through some visual display, where simple explanations are visible after the system performs an action, or through the vocalization of explanatory sentences as the system performs an action. Non-expert users shall not be required to expend additional effort to access relevant explanations. (see Gregor and Benbasat [B20] and Kulesza, Stupf, and Burnett [B35]). This interaction shall be adaptive to the user's interaction history as confidence is easily lost if e.g., the system behaves unexpectedly. Additional explanatory detail shall be available, on demand, as required by domain expert users or superusers, making it possible for them to interactively explore the system and its operation.

[a]Accessible means: in a format that is appropriate to the audio, visual or cognitive capabilities of the system's intended users.
[b]Information on references can be found in Clause 2.

5.1.2 The general public and bystanders

Transparency to wider society is needed in order to set expectations for the operation of autonomous systems and to help with building public confidence in the technology in an effort to reduce the potential for misuse and disuse of the technology. The role of the media in shaping public opinion is an important consideration here.

The general public are those persons who do not directly encounter an autonomous system but, nevertheless, may be affected directly or indirectly by its deployment. The public, through education, ethically aligned design in accordance with this and other standards, and legislation, should be empowered to make informed decisions if they want to become users and interact directly with an AIS. They should also understand the effects of the deployment of AI technology on their daily lives. However, it is well beyond the scope of this standard to discuss, let alone make suggestions on, societal concerns.

A subgroup of the general public are bystanders: persons who encounter an autonomous system without having any previous intention to achieve some purpose. This includes those simply observing the system function as well as those who may be passively impacted by it without their knowledge. For example, a person waiting at a train station where a mobile "customer help" system operates and serves another customer is a bystander as is someone entering a space which is monitored by a system using face recognition to identify occupants.

For this category of stakeholder, the levels of transparency are not progressive, i.e., fulfilment of an earlier level is not necessary to achieve a higher one with the exception of Level 3, which requires fulfilment of Level 2.

Transparency requirements for the general public and bystanders are given in Table 2.

Table 2—Transparency requirements for the general public and bystanders

Level	Definition
0 (lowest)	No transparency
1	The system shall be clearly identifiable by either a user or a bystander as an autonomous system. This requirement follows a proposed Turing Red Flag law: *An autonomous system should be designed so that it is unlikely to be mistaken for anything besides an autonomous system and should identify itself at the start of any interaction with another agent*. (Walsh [B58]). This identification shall be a simple message in the case of chatbots: a watermark on machine-generated multimedia, the use of stickers, or other insignia. Moreover, it may also be that a system design is structured in such a way that its manufactured nature is transparent (not anthropomorphic or zoomorphic, sensors are visible, etc.).
2	The system shall provide relevant warnings about any external sensor data collected or otherwise recorded (e.g., audiovisual input, geopositioning data, information gathered automatically) and which is related to the general public and bystanders. The system's manufacturer or operator shall provide documentation and/or identification graphics explaining what forms of sensor data are collected and how they are used, which shall be made publicly available. "Data which is related to general public and bystanders" refers to data from sensors in which the person is a feature. This level requires that the system's manufacturer or operator provides information on the types of data collected (i.e., metadata including, if applicable, personal data, but not the content of those data). See Level 4 for transparency of the data content. *The warnings may be physical cues on the robot and its environment, showing the location of sensors, similar to how body-worn cameras and CCTV require a sign to be present at the area of recording.* *The warnings may also be on-screen notifications, or a QR-style code, that leads to a source of further information about such sensors.* *Documentation may be leaflets containing all relevant information about the data used by the system(s). Another example is an autonomous vehicle manufacturer that provides online publicly accessible documents containing lists of sensors and explanatory data.*
3	All requirements of Level 2 shall be met. In addition, the documentation described in Level 2 shall also contain high-level descriptions of a system's intended purpose, a defined nominal operator of that system, as well as contact details for the system's owner, supervisor, or some other relevant authority where further information may be provided.
4	The system's responsible user shall have a clear data-governance policy and shall accept and respond to data-governance related requests. An example of such a data-governance policy is ISO/IEC 38505-1:2017 [B29] The system's owner may have an online form for data-governance requests, e.g., request of information stored. Once a person uses the form, the system owner receives the enquiry and processes it by returning an answer back to the requester.
5 (highest)	As Level 4.

5.2 Expert stakeholders who require transparency as part of their work

5.2.1 Validation and certification agencies and auditors

Software engineering distinguishes between verification and validation of software systems. This standard uses the term validation to encompass both these practices. It is important to note that this subclause does not require a system to have been verified or validated, instead it requires evidence of the verification and/or validation that has been undertaken, if any.

It is assumed here that many autonomous systems are subject to certification or evaluation processes in advance of deployment and in some cases at specific points in time after deployment in order to validate that the system is performing as desired. Such processes should be provided by agencies independent of the creators of the system. Certification might be a legal requirement (as, for instance, in the case of aircraft systems) or it might be a voluntary scheme providing some mark as a guarantor of quality. Similarly, assessments may be required by insurers and other bodies. The levels of transparency here can therefore be expected to correlate to the confidence such an agency can have in its determination of the quality of the system, though not necessarily any greater confidence in the quality of the system itself.

In general, certification, validation, and auditing are concerned with the *safety* and *data security* of a system, but there is no reason, in principle, why it should not also be concerned with qualities such as *reliability*, *robustness*, and so on. The levels of transparency are provided with this in mind.

Standards already exist for the validation of computational systems and a number of agencies already have reporting requirements, see for instance IEEE Std 1012 [B22] and ISO/IEC/IEEE 29119 [B31]. Nevertheless, autonomous systems present novel challenges to validation and are being deployed in situations where there is no obvious pre-existing regulatory body. This standard focuses on providing reporting requirements for the validation process of the whole autonomous system (that is, it focuses on the issue of the transparency of the validation of the autonomous system). Some of these requirements are relevant to any computational system, but some are of particular relevance to autonomous systems where the use of machine learning and embodiment are common. They may be used in conjunction with existing standards and processes and, indeed, the STS process outlined in Annex A may involve simply mapping existing reporting requirements to the appropriate transparency level. Where no pre-existing requirements exist then an STS must consider the appropriate level of transparency of the validation process for the application. The STS should note other transparency requirements in instances where there is not a perfect alignment between what a regulator demands and the content of this standard.

There are two aspects to the issue of transparency for validation and certification agencies. These two aspects are referred to as the *system description* and the *validation description*. In theory, such an agency should only require access to the full source code plus a physical example (in the case of a robotic system) of the system in order to be able to perform its own validation, but in practice it can be extremely difficult to understand how a system operates from only its source code (as an extreme example, it is currently impossible to adequately understand the functioning of a deep neural network after it has been trained). It is therefore expected that the work of validation and certification agents is best assisted by provision of details of any validation performed by the development team itself. In many cases the certification agency may be concerned primarily with certifying the company's *process* rather than validating the actual system. In most cases, it is assumed that the process includes ongoing validation of the system as it was developed.

The transparency levels in this standard assume that, in general, the more detail provided for one aspect, the more detail will be provided for the other. For instance, there would seem to be little point providing a reproducible validation artifact for a system that is only described by a specification. Therefore, these levels of transparency assume the minimum requirement of both aspects needing to be considered at that level.

In the transparency levels the primary concerns are with the following:

— *Specifications:* A specification is a description of what a system is intended to do (and not do). Without some sort of description of a system's purpose it is difficult for anyone to begin to make a determination about whether the system has any desired properties. While it is possible for specifications to be detailed, elaborate, and mathematical, this is not a requirement for transparency—there just needs to be some statement of purpose. Complex specifications frequently require validation processes of their own, for instance, to determine that they genuinely describe the system that is desired. Such validations of specifications may well be part of the validation process disclosed to the agency.

— *Properties:* The properties of a system can range from informal properties, such as "is easy to use," to precisely defined mathematical properties, such as "always applies the brakes within 0.5 s of detecting an obstacle." For an agency to begin to make a determination of the quality of a validation process, at a minimum they need to know what properties were considered.

— *Tests:* The field of testing computational systems is mature, and many techniques exist to help support testing that is appropriate and likely to catch important errors in a system, including testing for unintended outcomes of the system operation. So-called ad hoc testing, in which a developer or designer simply devises some relevant tests, is widespread and commonplace and may be sufficient for the validation of some properties of a system. Testing can also exist at several levels. System tests are applied to a completed system while unit tests are applied to components within a system. For some levels of transparency, only details of system tests are required and not the details of every test of every component.

— *Designs and models:* Nearly all complex systems have a high-level design. This may consist of documents outlining the main components of the system in natural language. However, a number of more formal notations exist for describing designs. Often, such formal notations comprise a mathematical model of the system itself that are executable in some fashion; the most obvious example would be a model of a robot in some simulated environment. While a system model or design does not provide every detail of the code, they often convey enough detail that testing and/or validation of the design/model allows major errors in the way the system operates to be detected. Provision of models and designs, therefore, allow the creator of an autonomous system to provide a great deal of useful information to an external agency while still protecting some intellectual property. ISO/IEC/IEEE 42010 [B32] and ISO/IEC/IEEE 42020 [B33] may be informative for architecture descriptions and architecture evaluation.

— *Statistical Models:* Many autonomous systems make use of statistical models derived from data to perform a range of tasks from situational awareness to full decision-making. Such models are created using a range of techniques including long standing statistical and optimization processes through to cutting-edge machine learning methods. The most well-known examples of such models are classifier systems used in image processing. These present challenges to validation. For instance, an object detection system's specification can often be no more precise than "identifies objects as reliably as most humans," and the classifier produced by the machine learning system may be difficult to understand even when full details of its operation are disclosed (often representing only statistical relationships between features of the system inputs). Many issues seen in such models arise from the data that was used to create them and to validate the performance of the system. There are well-documented cases of bias in such training data sets (e.g., sets of faces consisting primarily of young healthy people), leading to errors in system behavior and more general concern that a statistical model may have "blind spots" where no behavior has been learned for some combination of inputs. Therefore, higher levels of validation and certification transparency for autonomous systems that employ machine learning need access to training data, and access to the mechanisms by which the training data was assembled, in order to assess the risk of bias and omissions in the set. Data within machine learning systems may be in the form of a data set or encoded within models in the form of parameters or tokens. There are also wider concerns that such models may sacrifice fair or equitable behavior in preference for increasing the accuracy or optimality of some outcome. The validation of such models is a rapidly evolving field that includes purely technical advances in analyzing models with socio-technical techniques to assess

the impact of their deployment and understand the risks, particularly in terms of bias and fairness. This standard therefore focuses on transparency of the assessment process, explicit documentation of the risks as assessed, and any mitigations.

— *Source Code:* For the highest degrees of assurance of system behavior, an external agency may require access to the actual code of the system. At a minimum, this can allow such an agency to perform its own tests of the system, but a variety of techniques exist (including techniques based on mathematical proof) to assess the quality of a system based upon inspection of its code. Often, source code is difficult to understand for anyone except the programmer; this is one reason why it is important for agencies to have access to specifications and designs even when the source code is provided. This is why to achieve one level of transparency for validation and certification agencies and auditors, all the lower levels shall also be met.

— *Validation Tools:* Many tools exist to help with validation processes. These include tools for tracking the development process, the automated running of tests, running of tests on just parts of a system, mathematical validation of system models, assessments of the performance of machine learning, tools for analyzing the results of learning, and so on. Sometimes these tools may be proprietary and developed in-house at a particular company. For the highest level of transparency to be achieved where an agency is assessing a developer's validation process, executable versions of any tools used should be provided so that the agency can, if desired, reproduce the validation process.

For this category of stakeholder, the levels of transparency are progressive, i.e., fulfilment of an earlier level is necessary to achieve a higher one.

Transparency requirements for validation and certification agencies are given in Table 3.

Table 3—Transparency requirements for validation and certification agencies

Transparency level	Definition
0 (lowest)	No transparency
1	The system's developers shall provide documentation containing its specification and which of its properties were validated. *System Description:* A specification of the decisions to be taken by the system. *Validation Description:* A description of the validation process that was followed and which standards were applied.
2	The system's developers shall provide documentation containing its specification and description of its validation process. *System Description:* A specification of the system shall be supplied. *Validation Description:* A detailed description of the validation process shall be provided (including any ongoing validation processes used during system development or after deployment), including the specifics of system-level tests considered (where relevant). At this level and above, some internal validation (even if it is only ad hoc testing) shall have taken place. In addition to any general validation and verification information required by other certification processes, an analysis of the decisions to be made by the system and the validation of their implementation should be included.

Table continues

Table 3—Transparency requirements for validation and certification agencies *(continued)*

Transparency level	Definition
3	The system's developers shall provide documentation containing a high-level design or a (preferably executable) model of the system. The model may be a simulation of the final system. Statistical models used in the system should be documented along with the steps taken to validate their performance. If no models are used this should be explicitly stated. *System Description:* A high-level design or (preferably executable) model of the system shall be provided. This may be a simulation of the final system. *Validation Description:* An account of important issues uncovered and resolved during system development and/or deployment (as relevant at the time of submission) shall be provided even if full logs cannot be provided (e.g., because such logs are not kept). Where an analysis has taken place of the anticipated or actual operating conditions of the system (including unusual and hazardous situations) this should be provided. Where such an analysis has not taken place, this should be explicitly noted (with a justification, if desired). If statistical models are used by the system, an account shall be given of the steps taken to validate the performance of the model and the outcome of that validation. This account shall include discussion of any process undertaken to assess the possibility of unwanted bias, unfairness or inequity in the performance of the model, the outcome of that assessment, and steps taken to mitigate such issues (if any). The analysis of operating conditions should include any analysis of communities or environments that could be affected by the decisions of the system and the impact on those communities and environments, even where those communities and environments are not explicitly recognized as stakeholders. If no analysis of operating conditions has taken place and/or no assessment of statistical models has been made, this shall be stated. Full logs of any validation process of system decision-making should be provided if they exist, such as complete descriptions of test suites in terms of inputs provided and outputs observed, or outputs from proof tools (see Stepney and Polack [B54]). Any simulation model should itself be validated as providing a sufficiently high-fidelity model of the system and its environment, as relevant for its purposes, to allow its use in validation.
4	The system's developers shall provide a high-level design or (preferably executable) model of the system. This may be a simulation of the final system. Statistical models used in the system should be documented. If none are used, this should be explicitly stated. *System Description:* A high-level design or (preferably executable) model of the system shall be provided. This may be a simulation of the final system. Where relevant, all training data used in learning should be provided, including descriptions of the data's composition and provenance. *Validation Description:* All material necessary to reproduce the validation process for the final system shall be provided including, where relevant, executable versions of any tools used, and working versions of the system. In the case of a robotic system this should include a copy of the physical robot. Proprietary code may be provided in an executable form, provided the validation process remains reproducible. Where validation is being performed after deployment, the operational data collected as part of this validation should be provided. This shall include any analysis of the communities and environments affected by the system (whether intentionally or otherwise) and the effect observed. If no such analysis has taken place this shall be stated. It should be noted that for some systems it may be necessary for the certification/validation agency and developers to reach an agreement about data protection, and users may need to be informed about the use of their personal data for validation processes (including sharing it with external agencies).
5 (highest)	The system's developers shall provide the full source code, statistical models, and training data (if relevant), and any descriptions of the data composition and provenance. *System Description:* Full source code shall be provided, together with (where relevant) trained statistical models and all training data used in learning/optimization of those statistical models, including descriptions of the data's composition and provenance. *Validation Description:* All material necessary to reproduce the validation process shall be provided including, where relevant, executable versions of any tools used, and working versions of the system (including physical instantiations of the system where relevant).

It should be noted that this subclause is concerned only with the transparency of the validation process and the transparency of the system to external validators. The quality of the validation process is not of concern here; for example, whether specifications are well-constructed, appropriate properties are considered or the process is thorough.

While this subclause has concerned itself with transparency with respect to some particular agency, an autonomous system creator could choose to adopt these levels of transparency with regards to the general public, i.e., by placing system and validation descriptions somewhere publicly accessible where anyone could attempt to validate the system for themselves.

5.2.2 Incident investigators

If autonomous systems fail, they can cause a wide range of potential harm, from physical injury to psychological, economic, or environmental harm, thus processes for accident investigation are needed (see Winfield, Winkle, Webb, Lungs, Jirotka, and Macrae [B61]). This subclause of the standard defines the kinds and levels of transparency that support the work of accident (or more generally *incident*) investigators.

Failure of non-physical (i.e., software) systems can also cause harm. A medical diagnosis AI might, for instance, give the wrong diagnosis, or a credit-scoring AI might make a mistake 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.

An excellent model of good practice exists in the well-established and trusted processes of air accident investigation—processes that have contributed to the safety record of modern commercial air travel. Notably, air accident investigation agencies have a culture of learning and data sharing across the industry.

The ability to find out what went wrong and why is not only important to accident investigators; it might also be important in order to establish who is responsible, for insurance purposes, or in a court of law. In addition, following high profile accidents, wider society needs the reassurance of knowing that problems have been found and fixed.

The principle underlying this subclause is that, following an incident (which might have resulted in loss, harm or injury), it shall be possible to trace the internal processes of an autonomous system that, over some time period, led to the incident. This subclause requires that a system be equipped with a logging system for data, capable of securely recording a time-stamped log of key system inputs, outputs, and (ideally) high-level decisions (see Winfield and Jirotka, 2018 [B60]). In aviation, such devices are referred to as Flight Data Recorders, and in road vehicles they are known as Event Data Recorders. This standard adopts the term Event Data Recorder (EDR). The detailed specification of such an event data recorder is outside the scope of this standard, although for the specification of an EDR for motor vehicles refer to IEEE Std 1616-2021 [B23].

Incident Investigators are any persons or organizations tasked with discovering the root cause of an incident in order to make recommendations for corrective actions to prevent the future occurrence of the same or a similar event. Incident investigators normally have privileged (confidential) access to the system under investigation (or an identical copy should the system involved in the incident have been destroyed) together with designs and technical documentation. This standard expects that investigators also require access to information collected as a consequence of the transparency measures for safety certifiers set out in 5.2.1, in addition to the *event data* provided by the transparency levels defined in Table 4.

It is important to note that accident investigations are social processes of reconstruction that draw upon many sources of evidence including, for instance, eyewitness reports, CCTV, or other sources of video capture, forensic evidence, etc. Any information on the root cause of an incident collected through the transparency measures set out in Table 4 thus underpin and complement these other forms of evidence (see Winfield, Winkle, Webb, Lungs, Jirotka, and Macrae [B61]).

For this category of stakeholder, the levels of transparency are progressive, i.e., fulfilment of an earlier level is necessary to achieve a higher one.

Transparency requirements for incident investigators are given in Table 4.

Table 4—Transparency requirements for incident investigators

Transparency level	Definition
0 (lowest)	No transparency
1	A physical autonomous system such as a robot should be equipped with a video and audio recording device that is independent of the system's sensing and control systems and allows playback of the situation around the system at the time of an incident. The external data recorded by such a device should be relevant to the purpose and domain of application of the autonomous system, and such a device should be mounted appropriately, e.g., to face the direction of movement of an autonomous surface vehicle. The attribute of being "independent" of the system means that the device must be able to record unmodified, correctly time-stamped, and non-modifiable data through a means that is not dependent on the system itself, except for charging a battery on the device that provides a source of power independent from the system itself. A device such as a dashcam may suffice for these purposes. Software-only systems shall be equipped with an EDR module that logs both inputs to the system and outputs from the system.
2	Autonomous systems shall be equipped with an EDR capable of recording a time stamped log of key system inputs and outputs. A physical system such as a robot shall be fitted with an EDR capable of securely recording a time stamped log of key system inputs and outputs. The EDR's function is to continuously record the most recent *n* minutes or hours of relevant time-stamped data, including sensor data and actuator demands (as appropriate for the system in question). A physical EDR shall be designed and built to survive foreseeable accident and incident environments. Software-only systems shall be equipped with an EDR module that logs both inputs to the system and outputs from the system (as per Level 1).
3	Autonomous systems shall be equipped with an EDR designed to meet either a standard or open standard specification (where feasible standards exist), capable of recording a time stamped log of key system inputs, outputs and high-level decisions. A physical system, such as a robot, shall be fitted with either a physical or software EDR, as appropriate. The EDR's function is to continuously record the most recent *n* minutes or hours of relevant time-stamped data, including sensor data, actuator demands and high-level decisions (as appropriate for the system in question). These data shall be securely stored in a standard format. In the event that the physical system continues to function after the incident, the EDR shall continue recording after the incident. A physical EDR shall be designed and built to survive foreseeable accident and incident environments. Software-only systems shall have a standard or open standard (where feasible standards exists) EDR module that logs inputs to the system, outputs high-level decisions from the system in a secure, standard format.
4	The EDR in Level 3 shall additionally store the reason, e.g., decision-making logic or mechanism, behind each high-level decision, in order to allow an incident investigator to determine how and why the system made that decision. For autonomous systems in which decision-making is algorithmic, this requirement should be achieved by inserting calls to a procedure at each decision-making point in the code; each time that procedure is called, it sends a record identifying the decision-making point to the EDR. An incident investigator uses the trace of such decisions from the EDR, alongside inspection of the code, to determine the logic behind system decisions. For autonomous systems that make use of artificial neural networks (ANNs) the determination of the reasons for decisions (ANN outputs for a given set of inputs) is more difficult. But, at a minimum, the system should periodically send the complete set of ANN connection strengths to the EDR in order to allow incident investigators to reconstruct the ANN in an effort to reproduce the sequence of outputs leading up to the incident.
5 (highest)	In addition to the event data recorded to achieve Level 4, incident investigators shall be provided with a set of tools to assist them in reviewing and auditing that data. Such tools should provide visualization of the decisions made, e.g., in a tree-like format (see Theodorou, Wortham, and Bryson [B56]), or may even reconstruct a virtual model of the system.

5.2.3 Expert advisors in administrative actions or litigation

Designers of autonomous systems should be cognizant of the fact that agency administrative actions, lawsuits, or other legal proceedings may ensue when a system's operations directly or indirectly result in physical or economic harm. In such cases, the lawyers, judges, expert witnesses, and courts may require detailed information regarding how the system reached the state it was in when its operations resulted in harm. Without transparency, witnesses may be unable to provide an adequate description of the technology at issue or an adequate explanation of the specific system's actions, and the lawyers may not be able to adequately develop and present the evidence used in the legal process. Where factual evidence is not obtained through transparent investigations, the evidence could lead to unreliable conclusions, agency determinations, and court decisions that might harm public confidence in autonomous systems technology.

This standard expects that lawyers, judges, expert witnesses, insurers, or other professionals within this stakeholder group will require information collected as a consequence of the transparency measures set out in 5.2.1 and 5.2.2, i.e., the reports of both safety certification agencies and incident investigators, as the basis of their advice, judgements or testimony concerning a given system. If an incident involves human interaction with the system, they might also require information on transparency measures that are in place for the user, as set out in 5.1.1. However, it is expected that these professionals will also require evidence of the *processes* under which the system was designed, manufactured, or operated. These requirements for *process transparency* are set as follows:

— *Quality Management (QM)* is a process that seeks to help maintain consistency of an organization's product or service. QM has four main components: quality planning, quality assurance, quality control, and quality improvement (see Rose [B43]). QM is focused not only on product and service quality, but also on the means to achieve it.

— *Ethical Risk Assessment (ERA)* is a process that extends the envelope of risk assessment to include ethical risks. ERA assesses each risk of ethical harm, and the likelihood of that risk, then seeks ways of mitigating those risks. BS 8611:2016 [B9] provides guidelines for ethical risk assessment. IEEE Std 7000-2021 [B25] may also serve as a guide for this process.

— *Ethical Governance* is a set of processes, procedures, cultures, and values designed to help maintain the highest standards of behavior. Ethical governance thus goes beyond simply good (i.e., effective) governance, in that it inculcates ethical behaviors in both individual designers and the organizations in which they work (Winfield and Jirotka, 2018 [B60]).

— An *Audit Trail* is a chronological record, or set of records, that provides documentary evidence of an organization's processes. In the context of this standard, the audit trail shall document and record all quality, risk assessment and control/mitigation, and ethical governance processes.

For this category of stakeholder, the levels of transparency are non-progressive, i.e., fulfilment of an earlier level is not necessary to achieve a higher one.

Transparency requirements for expert advisors in administrative actions or litigation are given in Table 5.

Table 5—Transparency requirements for expert advisors in administrative actions or litigation

Transparency level	Definition
0 (lowest)	No transparency
1	Documentary evidence shall be provided to show transparent reporting of quality assurance activities for the system. Evidence of this may be demonstrated by the designer/manufacturer/operator of the system being conformant and certified to *quality management* standard ISO 9001:2015 [B27] or the equivalent.
2	The designer/manufacturer/operator shall undertake a process of *ethical risk assessment and control/mitigation* according to published standards such as BS 8611:2016 [B9], IEEE Std 7000-2021 Clause 11 (the section on transparency) [B25] or the equivalent and produce risk assessment reports for the system in question. ISO/IEC 33000 [B28] may provide guidance with regard to capability levels and process models for assessment. Such risk assessment reports shall detail which ethical risks were identified by the assessment, the likely impact of those risks, and the steps that have been taken to mitigate their impact.
3	In addition to Level 2, the designer/manufacturer of the system shall apply and document an *ethical governance* framework within its product life cycle. See for instance the 5 pillars of ethical governance set out in Winfield and Jirotka (2018) [B60].
4	For any given system there shall be a full *audit trail* for all of the quality, risk assessment and control/mitigation, and ethical governance processes in Levels 1–3 above. This audit trail may, for instance, form part of evidence within legal proceedings, internal investigations, or a public inquiry.
5 (highest)	As for Level 4.

Annex A

(informative)

A guide on how and when to use this standard

This standard has two primary functions. The first is as a tool for assessing the transparency of an autonomous system, and the second is as a guide to the transparency measures, for each stakeholder group, that should be taken into consideration during system specification and development. Note that this standard does not specify *how* the transparency measures defined here shall be implemented; only the kind of transparency each measure affords and how to determine whether it is present or not.

In this annex, an outline is provided on how to assess system transparency, then how to use this standard as a transparency design guide, and finally when to consider this standard.

A.1 How to assess system transparency

Each of the definitions for the different levels of transparency set out in Clause 5 is a testable specification that, for any given system, will either be met or not met. Overall system transparency is therefore assessed by working through each transparency level definition, for each stakeholder group in turn, to answer the yes/no question “Does the system meet this transparency level specification or not?” The STA checklist in Table A.1 can assist in this process.

Table A.1—STA template

STA **System:** **Assessor:** **Date:**			
Standard Clause	**Level**	**Yes/No**	**Notes**
5.1.1 Users	1		
	2		
	3		
	4		
	5		
5.1.2 General public and bystanders	1		
	2		
	3		
	4		
5.2.1 Validation certification agencies and auditors	1		
	2		
	3		
	4		
	5		
5.2.2 Incident investigators	1		
	2		
	3		
	4		
	5		
5.2.3 Expert advisors in administrative actions or litigation	1		
	2		
	3		
	4		

The overall transparency assessment is summarized using Table A.2.

Table A.2—STA scoresheet

STA Scoresheet **System:** **Assessor:** **Date:**					
Standard Clause **(C = cumulative, NC = non cumulative)**	**Levels** **(tick to indicate level is met)**				
	1	**2**	**3**	**4**	**5**
5.1.1 Users (NC)					
5.1.2 General public and bystanders (NC)					
5.2.1 Validation and certification agencies (C)					
5.2.2 Incident investigators (C)					
5.2.3 Expert advisors in administrative actions or litigation (NC)					

A.2 How to use this standard as a transparency design guide

There are many reasons a designer might consider designing transparency into a system. These include the following:

a) The system has the potential to cause harm, noting that harms could be physical, psychological, economic, societal, or environmental.

b) The system might capture personal information (and make decisions or recommendations based on that personal data) and therefore be subject to data protection regulations such as General Data Protection Regulation (GDPR).

c) The user should have confidence in the system; for instance, the success of the system could depend on a (possibly non-expert) user having high confidence in that system, and in order to build that confidence the user needs to gain a good understanding of what the system does, why it does it, and when.

d) The system will be deployed in publicly accessible buildings (e.g., shopping malls, hospitals, or museums) or urban spaces (e.g., streets or public parks). Users of those spaces who do not interact directly with the system (e.g., pedestrians, shoppers, families and children, public servants including police, paramedics or street cleaners) may require some understanding of what the system is and what it does.

e) The customer for the system (which might be a government department) writes the need for transparency into the System Requirements Specification and makes the award of a design contract subject to conformance with those transparency requirements.

f) The system design company is committed to practicing Ethically Aligned Design within a broader framework of Responsible Innovation and regards transparency as an important design principle for its products and services.

This standard has an important role as a guide for system procurers or designers who for any reason, including those outlined above, are considering which transparency features need to be incorporated into the system specification. An outline process for preparing an STS is shown in Figure A.1.

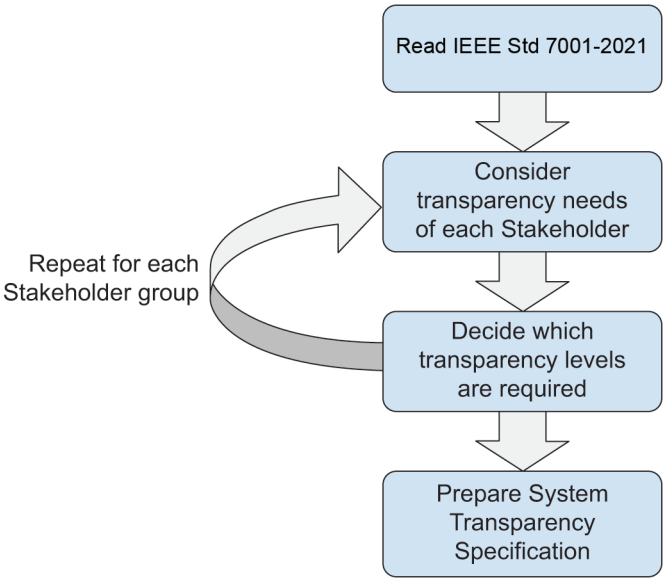

Figure A.1—Outline process for preparing an STS

Each of the four main steps in the outline process of Figure A.1 are detailed as follows:

— *Step 1:* Read this standard. Before starting the process of drafting the STS, it is important to understand the overall transparency framework set out in this standard—especially the need to think about transparency needs from the perspective of the five stakeholder groups.

— *Step 2:* Consider the transparency needs of each stakeholder group as set out in Clause 5. Each system will have different transparency priorities, as will various stakeholders alike. As outlined previously in transparency design considerations a) through f), these might be transparency for: minimizing harm, data protection, improving user confidence, to meet customer requirements, or as part of Ethically Aligned Design.

— *Step 3:* Decide which transparency levels are required. Not all systems will need to meet the maximum levels of transparency defined in Clause 5, and the balance of transparency needs will vary across stakeholder groups given the transparency priorities that apply to the system and its application under consideration. The decision of which transparency level is required for each stakeholder group should be made following an impact analysis. That impact could, for instance, be classed as high, medium, or low. Safety-critical autonomous systems, which have the potential to cause serious harm or injury, would be classed as high impact. Recommender systems (AIs that do not make decisions directly but instead support a human decision maker) might be classed as medium impact, while systems with little or no real-world consequence would be classed as low impact. High impact systems would then require greater transparency than medium impact, which in turn would require greater transparency than low impact systems. It should be noted that these impact assessments are independent across stakeholder groups, so a high impact for one group does not necessarily imply a high impact across all groups. Analysis may be required to explore the relative impact of transparency or explainability decisions for various groups of stakeholders. For example, the meaning of greater transparency for a high-impact system, such as an autonomous aircraft (drone), to bystanders, users, system owners, designers, and forensic analysts is quite different because of their level of understanding, ability to influence the system, and the likelihood of being affected by system hazards. Greater transparency for high impact is therefore not a one-size-fits-all requirement. The scenarios included in Annex B are intended to illustrate how transparency is either measured or specified in different fictional applications and situations.

— *Step 4:* Prepare the STS. After repeating Step 2 and Step 3 for each stakeholder group, the STS can be drafted. An STS template is given in Table A.3.

Table A.3—An STS template

STS Template **System:** **Specifier:** **Date:** **Notes on overall transparency priorities:**					
Standard subclause **(C = cumulative, NC = non cumulative)**	**Levels** **(tick to indicate levels required)**				
	1	**2**	**3**	**4**	**5**
5.1.1 Users (NC)					
Notes:					
5.1.2 General public and bystanders (NC)					
Notes:					
5.2.1 Validation and certification agencies (C)					
Notes:					
5.2.2 Incident investigators (C)					
Notes					
5.2.3 Expert advisors in administrative actions or litigation (NC)					
Notes:					

A.3 When to apply this standard

How this standard is best applied depends upon when in the development lifecycle the standard is taken into consideration. This standard may be applied at any stage, from requirements specification (as outlined in A.2), then at any stage during development and deployment. Given that transparency does not come "for free," but needs to be designed-in, then the greatest benefit (at the lowest cost) can be gained from this standard by considering transparency early during the development lifecycle—the earlier the better.

Consider now how to apply this standard at different stages in a system life cycle, as follows:

— *During system specification:* This standard can be employed during the requirements specification phase in order to consider and prioritize transparency needs, then prepare an STS, as detailed in A.2. The STS then becomes part of the overall System Requirements Specification against which design can proceed.

— *During design and development:* Although the process of STA outlined in A.1 may be applied at any time during the system development phase, early application of STA is clearly advantageous as it enables any transparency deficits to be addressed during initial builds of the system.

— During system deployment. System transparency may be assessed (using the method in A.1) while a system is in use. This may be valuable to, for instance, compare the transparency of different systems or, following a system failure, to retrospectively assess its transparency in order to learn lessons for future systems.

It is important to note that the application of this standard during the design and deployment life cycles should not be a one-off process. Instead, this standard should be applied iteratively, for instance following major system revisions, or following a change in the way the system is deployed. Thus, this standard can be used to check and demonstrate that system updates or operational changes have not resulted in either reduced transparency or transparency that is no longer sufficient.

Annex B

(informative)

Scenarios

B.1 Autonomous delivery vehicle

This fictional scenario illustrates the value of conducting a STA early in the development process.

An established and well-regarded manufacturer of robots for indoor use, including hospital portering robots, wishes to expand its range into Autonomous Vehicles designed to provide delivery services between local suppliers and their customers, including deliveries of both groceries and hot food.

The company has built a demonstrator system. Early in the design cycle they conduct a series of real-world trials involving a number of local suppliers including a local supermarket and two fast food outlets, and a panel of volunteer customers. The manufacturer regards themselves as a responsible company who fully understands that, to be successful, the delivery autonomous vehicle will need to be both reliable and have a low risk of causing harm. They conduct a STA against this standard, with the aim of considering the STA alongside feedback from the real-world trials. The score sheet summarizing the outcomes of that assessment is shown in Table B.1.

Table B.1—Autonomous delivery vehicle STA scoresheet

System: Autonomous delivery vehicle **Assessor: Dr J Bloggs** **Date: 23 March 2021**					
IEEE Std 7001-2021 subclause (C = cumulative, NC = non cumulative)	**Levels (tick to indicate level is met)**				
	1	**2**	**3**	**4**	**5**
5.1.1 Users (NC)	X * ***	X ** ***	X ***		
NOTE—Three categories of users are defines as follows: *Customers who have placed an order and need to interact with the autonomous vehicle in order to collect their food delivery. For these users simple instructions, with images and a video-clip explaining how to collect the delivery from the autonomous vehicle are provided when an order has been accepted and delivery confirmed. Pictorial instructions are clearly displayed on the vehicle and, in addition, spoken instructions are triggered when the person collecting the order approaches the AV: Level 1 **Non-expert persons responsible for placing the order into the autonomous vehicle prior to sending it out for delivery. For these interactive training materials are provided: Level 2. ***Domain expert users are defined here as the operators of the AV, who will monitor and supervise its operation and, when necessary, maintain the vehicle. Domain experts are provided with full technical documentation (Level 1), together with interactive training materials (Level 2) and functionality to provide a full explanation of the AVs activity (Level 3).					
5.1.2 General public and bystanders (NC)	X	X			
NOTE—The autonomous vehicle is clearly identified as a robot, with warnings; it is fitted with cameras for navigation, with limited views such that they do not collect personal data.					
5.2.1 Validation and certification agencies (C)	X	X			
NOTE—Transparency of validation processes up to Level 2.					
5.2.2 Incident investigators (C)	X	X			
NOTE—The present system is equipped with a proprietary event logging system.					
5.2.3 Expert advisors in administrative actions or litigation (NC)	X	X			

NOTE—The company has ISO 9001 [B27] accreditation or equivalent and ethical risk assessment (ERA) has been undertaken for the AV.

When reviewing the STA the company notes that the transparency measures for non-expert users reflect the satisfaction with the information provided that was reported by them following the trials. However, the company noted that they had not yet conducted trials with a potential third-party operator and therefore could not be confident that the transparency measures for domain expert users are sufficient. By the same token the STA prompted the company to conduct trials with a range of bystanders in order to determine whether the measures in 5.1.2 are considered sufficient.

For section 5.2.2 of the STA: incident investigators, the company, and its insurers decide that Level 2 is not sufficient as the current proprietary event data recorder fitted does not record the reasons for the AVs decisions. Given that the autonomous vehicle will be operating in public spaces, safety is paramount. Thus, the ability to fully investigate both near-miss and actual accidents will be essential in improving both the AVs safety features and operational processes.

B.2 Medical diagnosis AI

This fictional scenario shows how this standard can be used to specify system transparency requirements as a condition of supply.

A government procurer of health technology believes that clinicians (both in general practice and in hospitals) would benefit from an AI-based tool to assist them in reaching diagnoses. Based upon a good understanding of the state of the art in diagnostic AI systems they write a specification for a Medical Diagnosis AI Recommender system and decide that the system should meet or exceed necessary levels of transparency as a condition of supply. Using this standard as a guide, they draft the following STS for the recommender, for inclusion in the call for tenders.

Table B.2—Medical diagnosis AI system transparency scoresheet

System: Medical diagnosis AI (recommender) system **Specifier: Government Department of Health** **Date: 24 September 2021** **Notes on overall transparency priorities: The recommender system requires a high level of transparency for both its recommendations to a clinician, and for the processes used to develop the system and to validate its operation.**					
IEEE Std 7001-2021 subclause (C = cumulative, NC = non cumulative)	**Levels (tick to indicate level is met)**				
	1	**2**	**3**	**4**	**5**
5.1.1 Users (NC)	X	X	X	X	
NOTE—Users are defined as clinicians in the category of domain expert users, who require a high level of understanding of how the recommender system functions, including the ability to ask it to explain its recommendations.					
5.1.2 General public and bystanders (NC)	X	X			
NOTE—This stakeholder group is less critical, since the clinician is required to explain to a patient (and family members, etc.) the role and purpose of the recommender system in helping to reach a diagnosis.					
5.2.1 Validation and certification agencies (C)	X	X	X	X	
NOTE—Evidence of validation, including clinical trials is critical.					
5.2.2 Incident investigators (C)	X	X	X	X	
NOTE—The recommender system must securely log all recommendations, including the reasons for those recommendations, to support incident investigations, noting that an incident investigation may be triggered by a clinician raising concerns about the system's recommendations.					
5.2.3 Expert advisors in administrative actions or litigation (NC)	X	X	X	X	X

NOTE—The fullest possible evidence of best practice quality management, development, and governance processes in the supplier is required.

The Department of Health includes the STS (Table B.2) in the call for tenders for the recommender system. The call requires suppliers to demonstrate compliance by detailing the following in their bids:

a) How the transparency measures required have been implemented

b) IEEE 7001 STAs that clearly show that the transparency measures meet or exceed the specifications in IEEE Std 7001-2021 (Table B.2).

B.3 Content moderation for AI

This fictional scenario shows how this standard can be used to specify system transparency requirements as a condition of supply.

A video hosting website has been accused by activists of using keywords to prevent monetization of potentially objectionable or controversial content. The activists attempted to reverse-engineer the algorithm and have created a list of keywords that they believe can trigger the content moderation algorithms to demonetize content. To mitigate a potential scandal and lawsuits, and to satisfy legislators, the video hosting website decides to apply this standard on transparency as a draft specification for their engineers, to more transparently communicate the decision-making processes of their content moderation systems.

Table B.3—Content moderation for AI system transparency scoresheet

System: Content moderation AI system **Specifier: Video hosting website** **Date: 11th November 2021** **Notes on overall transparency priorities: The Content Moderation AI System requires a high degree of transparency for legislators and auditors, but with less transparency for the general public, due to concerns of bad actors finding exploits.**					
IEEE Std 7001-2021 subclause (C = cumulative, NC = non cumulative)	**Levels (tick to indicate level is met)**				
	1	**2**	**3**	**4**	**5**
5.1.1 Users (NC)	X	X	X		
NOTE—Users are defined as content creators, who are non-expert users. They require a medium level of understanding of how the system functions, including the ability to ask the system to explain its decisions, or to preemptivelyinterrogate if something is likely to be deemed problematic.					
5.1.2 General public and bystanders (NC)	X				
NOTE—This stakeholder group is defined as content consumers, who are only indirectly affected by potential issues relating to content moderation and related monetization.					
5.2.1 Validation and certification agencies (C)	X	X	X		
NOTE—Evidence of validation of the algorithm is important for illustrating good faith, and they require a medium range of information.					
5.2.2 Incident investigators (C)	X	X	X	X	
NOTE—Incident investigators and auditors should have privileged access to the mechanisms, in order to better ascertainif they are fair and appropriate or are harming any interests unfairly.					
5.2.3 Expert advisors in administrative actions or litigation (NC)	X	X	X	X	X
NOTE—Legal and legislative concerns may demand, or subpoena confidential information related to the system in the line of their duties.					

The video hosting website includes the STS (Table B.3) in the specification for the content moderation AI system. The specification requires engineers to demonstrate compliance by detailing, in their bids a) how the transparency measures required can be implemented and b) IEEE 7001 STAs that clearly show that the transparency measures meet or exceed the IEEE 7001 specification (Table B.3).

B.4 Credit scoring system

This fictional scenario shows how this standard can be used to specify system transparency requirements as a condition of supply.

A credit scoring technology wishes to illustrate to loan applicants, service users, and legislators that their technologies are open and safe. The credit scoring company decides to apply this standard on transparency as a draft specification for their engineers to more transparently communicate the decision-making processes of their content moderation systems.

Table B.4—Credit scoring system transparency scoresheet

System: Credit scoring system **Specifier: Loans company** **Date: 11th November 2021** **Notes on overall transparency priorities: The Credit Scoring System requires a high degree of transparency for legislators and auditors, but with less transparency for the general public, due to concerns of bad actorsgaming the system.**					
IEEE Std 7001-2021 subclause (C = cumulative, NC = non cumulative)	**Levels (tick to indicate level is met)**				
	1	**2**	**3**	**4**	**5**
5.1.1 Users (NC)	X *	X * **	X *		
NOTE—Two categories of user are defines as follows: *Loan applicants, who are non-expert users. Transparency is very important to this group as the assessment is of their own particulars, and they deserve a chance to understand why they have been assessed in a particular way, and to seek redress in the event that information is incorrect or is assessed unfairly. **Operators of the credit scoring system who are assessing potential clients for creditworthiness, which may also be applied as a proxy for trust in scenarios not related to credit per se. These are expert domain users.					
5.1.2 General public and bystanders (NC)	X				
NOTE—The system requires there to be less transparency for the general public due to concerns of bad actors gaming the system.					
5.2.1 Validation and certification agencies (C)	X	X	X	X	
NOTE—Evidence of validation and certification to a high degree is essential, given the sensitivity of the system.					
5.2.2 Incident investigators (C)	X	X	X	X	X
NOTE—The credit scoring system must securely log all recommendations, including the reasons for those recommendations, to support incident investigations, noting that an incident investigation may be triggered by an operator, watchdog, or ombudsman raising concerns about CSS's recommendations.					
5.2.3 Expert advisors in administrative actions or litigation (NC)	X	X	X	X	X

NOTE—Legislators should have highly privileged access to information, as loss of economic franchise based on a protected characteristic may be unlawful.

The loans company includes the STS (Table B.4) in the specification for the credit scoring system. The specification requires engineers to demonstrate compliance by detailing, in their bids a) show the transparency measures required can be implemented and b) IEEE 7001 STAs that clearly show that the transparency measures meet or exceed the IEEE 7001 specification in Table B.4.

B.5 Security robot

This fictional scenario shows how this standard can be used to specify system transparency requirements as a condition of supply.

A security company wishes to deploy a new security robot system that must prioritize public safety without being easily exploitable or gamed. The security company decides to use this standard on transparency as a

draft specification for their engineers to more transparently communicate the decision-making processes of their security systems.

Table B.5—Security robot system transparency scoresheet

System: Security robot **Specifier: Security company** **Date: 11th November 2021** **Notes on overall transparency priorities: The security robot requires a high degree of transparency for legislators and auditors, but with less transparency for the general public, due to concerns of criminals finding exploits.**					
IEEE Std 7001-2021 subclause (C = cumulative, NC = non cumulative)	**Levels (tick to indicate level is met)**				
	1	**2**	**3**	**4**	**5**
5.1.1 Users (NC)	X	X	X		
NOTE—Users are defined as deployers and administrators of the security robot, who may be site managers, or who may be a third-party contractor. They require a medium level of understanding of how the guard bot functions, including the ability to ask it to explain its protocols or predict its behavior in a given situation, and repair simple faults. These are superusers.					
5.1.2 General public and bystanders (NC)	X	X			
NOTE—This stakeholder group is important to bear in mind for matters of public safety, though this group is potentially adversarial, and so warrants less disclosure.					
5.2.1 Validation and certification agencies (C)	X	X	X	X	
NOTE—Electromechanical devices that could potentially cause serious injury warrant a high degree of certification and oversight.					
5.2.2 Incident investigators (C)	X	X	X	X	X
NOTE—The security robot securely logs all actions and behavior of self, and other agencies in the vicinity. With regards to the behavior of the systems itself, it should log the reasons why it made a certain appraisal, prediction, decision, or action. Investigation may be called in the case of an altercation causing alarm and distress or injury.					
5.2.3 Expert advisors in administrative actions or litigation (NC)	X	X	X	X	X

NOTE—Legal and legislative concerns may demand, or subpoena confidential information related to the robot in the line of their duties.

The security company includes the STS (Table B.5) in the specification for the security robot. The specification requires engineers to demonstrate compliance by detailing in their bids a) how the transparency measures required can be implemented and b) IEEE 7001 STAs that clearly show that the transparency measures meet or exceed the IEEE 7001 specification above (Table B.5).

B.6 Medical decision support system

This fictional scenario shows how this standard can be used to assess system transparency in two similar systems in a similar context, in this case that of a medical decision support system (Med DSS).

This scenario is focused on a Med DSS that uses a machine learning (ML) algorithm to provide recommendations regarding who should receive a kidney transplant within a group of compatible patients. Two cases vary in the degree of automation complexity and human oversight. In both cases, if a wrong decision is made, there may be severe consequences for the patient and others who might have received the organ transplant. Thus, the decision is characterized by high criticality.

In both cases, the training data set used for initially training the model came from patients aged 18 to 35 enrolled in an NHS trial in the UK. Thus, the DSS recommendation might be biased. Additionally, in both cases, the DSS uses the following profiling approach: a particular gene complex is associated with better outcomes. The system finds associated genotypic factors and uses this in decision making.

B.6.1 System Version 1

In the first case, the DSS uses an algorithm that is comprehensible to developers of the algorithm but not the end-users. The DSS uses specific data inputs known to influence kidney transplant success rates (e.g., age, hospital facilities, and distance to a donor) to make a recommendation. There is a significant oversight by humans on the performance of the DSS. The DSS is acting as part of a team with human consultants/clinicians who provide specialist expertise. Where traditional processes would refer the decision to a team of five clinicians, the decision is now made by four clinicians and the recommendations of the algorithm. The algorithm uses patterns based on the training data and is not provided with additional information. Table B.6 is a worked example of the transparency assessment of this system.

Table B.6—Med DSS Version 1 system transparency scoresheet

STA **System: Med DSS** **Assessor:** **Date:**			
Standard subclause	**Level**	**Yes/No**	**Notes**
5.1.1 Users *Users in this case are the hospital clinicians involved in the kidney transplant process (domain expert users)*	1	Yes	The users are provided with documentation including general principles of operation and the source of the training data set
	2	Yes	Clinicians have available interactive training material to rehearse interactions
	3	Yes	Clinicians can query the system to receive an explanation of recent activity
	4	Yes	Clinicians can receive information on what the system *would* do in a given situation
	5	Yes	Clinicians are provided with continuous on-demand explanation of behavior. However, this does *not* include access to training data because this contains sensitive medical information
5.1.2 General public and bystanders	1	Yes	The general public (including patients) are aware that an AI is a member of the clinical team
	2	No	No information is given to patients on data collected
	3	No	No information is given to patients on the system purpose, goes and operation
	4	No	There is no data-governance policy
5.2.1 Validation and certification agencies and auditors	1	Yes	Documentation containing specification and which of its properties were validated is available
	2	Yes	Documentation containing validation processes is available
	3	Yes	Documentation containing a high-level design of the system is available including composition and provenance of training data
	4	Yes	Documentation containing a high-level design of the system is available including composition and provenance of training data
	5	Yes	The full source code including profiling information is available.
5.2.2 Incident investigators	1	Yes	The system logs both inputs and outputs of the system
	2	Yes	The system logs both inputs and outputs of the system
	3	Yes	The system logs inputs, outputs and high-level decisions in a secure, standard format
	4	Yes	As Level 3 with the addition of the likely reasons for the decisions
	5	Yes	As Level 4 with the addition of tools to audit the data
5.2.3 Expert advisors in administrative actions or litigation	1	Yes	Documentary evidence of quality management standard compliance is available
	2	Yes	An explicit process of ethical risk assessment and control/mitigation has been undertaken
	3	No	The designer/manufacturer of the system did not apply a documented and transparent ethical governance framework
	4	No	An audit trail is not present

B.6.2 System Version 2

In the second case, the Med DSS uses an algorithm that is not easily comprehensible to domain expert end-users. The learning algorithm collects large volumes of patient data including biometrics, longitudinal health information for that patient, and other kidney recipients not normally accessible to the clinicians. The system processes this information to deliver recommendations. Clinicians define a list of 10 compatible patients and then the algorithm makes the selection of which patient from that list receives the transplant. There is limited/minor oversight by humans on the performance of the DSS. Decisions are reviewed regularly to validate the process. The algorithm uses patterns based on training data and additional information on the prevalence of this gene across different ethnicities. Table B.7 shows the information from the second case described above:

Table B.7—Med DSS Version 2 system transparency scoresheet

STA **System: Med DSS** **Assessor:** **Date:**			
Standard subclause	**Level**	**Yes/No**	**Notes**
5.1.1 Users *Users in this case are the hospital clinicians involved in the kidney transplant process (domain expert users)*	1	Yes	The users are provided with documentation including general principles of operation and the source of the training data set.
	2	Yes	Clinicians have available interactive training material to rehearse interactions
	3	No	It is not possible to receive a brief and immediate explanation of the deep learning algorithm's decision-making process
	4	No	It is not possible to receive a brief and immediate explanation of the deep learning algorithm's decision-making process
	5	No	Clinicians are not provided with continuous on-demand explanation of behavior or access to training data that contains sensitive medical information
5.1.2 General public and bystanders	1	No	The general public will not be aware of this system.
	2	No	No information is given to patients on data collected.
	3	No	No information is given to patients on the system purpose, goals, and operation.
	4	No	There is no data-governance policy.
5.2.1 Validation and certification agencies and auditors	1	Yes	Documentation containing specification and which of its properties werevalidated is available
	2	Yes	Documentation containing validation processes is available
	3	Yes	Documentation containing a high-level design of the system is available including composition and provenance of training data
	4	Yes	Documentation containing a high-level design of the system is available including composition and provenance of training data
	5	Yes	The full source code including profiling information is available.
5.2.2 Incident investigators	1	Yes	The system logs both inputs and outputs of the system
	2	Yes	The system logs both inputs and outputs of the system
	3	Yes	The system logs inputs, outputs and high-level decisions in a secure, standard format
	4	No	Reasons for the decisions are not available
	5	No	No tools to audit the data are available
5.2.3 Expert advisors in administrative actions or litigation	1	Yes	Documentary evidence of quality management standard compliance is available
	2	Yes	An explicit process of ethical risk assessment and control/mitigation has been undertaken
	3	Yes	The designer/manufacturer of the system applied a documented and transparent ethical governance framework.
	4	No	An audit trail is not present.

B.6.2.1 Overall transparency assessment

The overall transparency assessment is summarized using the Table B.8.

Table B.8—Overall system transparency scoresheet

STA Scoresheet **System:** **Assessor:** **Date:**					
IEEE Std 7001-2021 subclause (C = cumulative, NC = non cumulative)	**Levels (tick to indicate level is met)**				
	1	**2**	**3**	**4**	**5**
5.1.1 Users (NC) (System 1)	X	X	X	X	X
5.1.1 Users (NC) (System 2)	X	X			
5.1.2 General public and bystanders (System 1)	X				
5.1.2 General public and bystanders (System 2)					
5.2.1 Validation certification agencies (C) (System 1)	X	X	X	X	X
5.2.1 Validation and certification agencies (C) (System 2)	X	X	X	X	X
5.2.2 Incident investigators (C) System 2)	X	X	X	X	X
5.2.2 Incident investigators (C) (System 2)	X	X	X		
5.2.3 Expert advisors in administrative actions or litigation (System 1)	X	X			
5.2.3 Expert advisors in administrative actions or litigation (System 2)	X	X	X		

B.7 Increasing levels of mainline railway automation

B.7.1 Background

This fictional scenario based on a real context shows, via a rail example, the need for transparency and how this need grows as operation moves from automated with human oversight (human delegated or supervised) to fully autonomous.

Rail systems already have significant levels of automation, right up to what is known as Unattended Train Operation [(UTO), Grade of Automation level 4 (GoA4) under the IEC Standard for Communications Based Train Control (CBTC) (see Schifers and Hans [B48])] where all functions, including door operation, are performed by the control system and there is no crew on the train at all, even for emergencies. Such systems are already common on metros and people movers. These are normally “closed” (the route is largely in a tunnel or elevated) where they are at ground level and they are protected by substantial fences. Train speeds are relatively low. There are no level crossings of any kind, and platforms are often protected by platform screen doors (PSDs) such that all access to the track/guide-way is controlled. The systems are usually geographically constrained, so emergency and recovery response can be provided in a timely manner from off-route resources. These systems are currently based on validated software produced by conventional programming.

Increasingly there is a desire to apply automation to mainline railways, and it can be difficult to explain to non-railway people why this is much harder.

One of rail's competitive advantages is that a steel wheel on a steel rail has a very low rolling resistance and therefore trains are generally energy efficient if effectively loaded.

The downside of this feature is that this also leads to low available friction, particularly if the rails are wet and or contaminated, meaning very long stopping distances. As a result, at any significant speed, control systems are required that give drivers and/or automatic driving systems information about the status of things

happening beyond visual range. Protection from conflicting routes is provided by interlocking systems based on highly validated conventional algorithms (often using formal methods). Driver/system observance of route commands (signals, where lineside signaling is retained) is enforced by Automatic Train Protection (ATP), a highly validated computer system based on calculated braking curves and limited movement authorities (generated by validated algorithms).

Hazards that the driver can see may lead to some control action (principally application of the brakes, since there is no steering) or issuance of a warning by sounding the horn but the effect of this may only be mitigation rather than prevention and there are circumstances where such action could make matters worse.

Mainline railways are rarely closed; indeed only a few railways have a duty to fence the track (the UK being one) and, apart from on modern high-speed lines, there are often both vehicle and footpath level (at grade) crossings. The greatest level of harm on most mainline railways is trespass and suicide, followed by level-crossing collisions/incidents. Despite being "open" systems, access to the track for emergency vehicles can be quite challenging with long track lengths through rural areas.

Recognizing these issues, there is a current tendency to propose solutions based on Automatic Train Operation (ATO), where the system drives the train but a human is retained in the cab for monitoring and secondary safety. It is further suggested that as the LIDAR and imaging technologies being developed for autonomous road vehicles mature, the human driver will be able to be replaced. This leads to a number of both technical and ethical issues and drives a need for high levels of transparency in any AI system developed and deployed.

For instance, if a person is in close proximity to the track, a very finely nuanced decision may be required to predict that person's intent and level of concentration. Are they distracted (for instance looking at their phone with their earphones in)? Are they moving at a pace where they will likely come into conflict or likely be clear before the train arrives? Do they have anything with them that could cause additional issues like a baby buggy (stroller). Have they seen the train and are clearly waiting for it to pass or are they looking at the train and moving in a way that might mean they are contemplating suicide? All of that evaluation has to be achieved in a very limited time period against a high level of near-field and far-field clutter and under all lighting conditions, including in the dark with headlights.

Drivers get a feel for such things, and it may be challenging to build and operate a self-learning system that can mirror that. While some of these detection issues are common to autonomous road vehicles the inability to stop before any item of interest introduces a very different kind of complexity. Trains cannot steer away; the only control available to the driver or the automatic system is the brake. If the emergency brake is applied, under current design philosophies the ATO will disengage, and a number or conditions need to be met to re-engage it. Thus, a high false alarm rate would potentially be very disruptive without a design change to the ATO philosophy, which would have other implications. On the other hand, application of the brakes might avoid or substantially mitigate a potential accident. But braking if a vehicle approaching a crossing looks like it is not going to stop might encourage a road driver to gamble, and a train hitting a car can be much more damaging than a car hitting the barrier or even the side of a train. Additionally, sounding the warning horn too often may be considered a noise nuisance and create an adverse reaction, particularly if the need for the warnings cannot be fully explained and justified. Balancing the need for early detection with a low false alarm rate will be very challenging. Further, glancing blows with people or animals are relatively common, and these might be quite hard to detect. Someone being found injured (or having died of their injuries near the lineside) sometime after the event without any warning flag could cause a significant public outcry and require a detailed independent investigation that would expect and could be aided by transparency regarding the sensor data, the system's resultant actions, and, where appropriate, interaction with human drivers/operators.

So, transparency in what the system sensors saw and the resulting decisions will be essential in investigating any accidents or incidents on a regular basis, not just occasionally. Such information will also need to be stored in a secure manner for at least several days as it may not be immediately apparent that an incident has occurred. New routes will have specific features that will have to be learned and accommodated and, if there is an

incident, it will be very important to understand whether a wrong decision was made or whether the situation was simply unavoidable.

The scene should also recognize the potential upsides in that this task needs to be performed in all weather and at night, so a number of available sensors may offer a potential improvement in detection accuracy over the human eye aided only by headlights.

B.7.2 Two cases in which this standard might be used

B.7.2.1 AI assists the driver

An AI system provides assistance to a human train supervisor who makes the final decision as to whether to apply the brakes, sound the horn, or report an incident to 'control.' This case has three potential sub-cases, as follows:

a) The AI system does not generate an alert, nor does the human operator see anything, but an incident occurs.

b) The AI system generates an alert, but the human operator does not heed it, and an incident occurs.

c) The AI system generates an alert, and the human operator responds to it in a timely manner, but an incident still occurs.

In each of these cases transparency will be needed to understand why the system responded (or did not respond) in the way it did. Cases b) and c) will have related sub-cases where there was an alert, but an incident was avoided. In these cases, while there may be no pressing need for investigation, transparency may still be required to support performance improvement. So, consider how this standard can be applied in this case (Table B.9).

Table B.9—AI assistance system transparency scoresheet

System: ATO with AI driver assistance **Assessor: A Safety Engineer** **Date: xx.xx.xx**			
IEEE Std 7001-2021 subclause	**Level**	**Yes/No**	**Notes**
5.1.1 Users: *Drivers (domain expert users) and their train operating company employers (safety duty holder) (mix of domain expert and superusers), train owners, train builders (superusers)*	1	Y	All
	2	Y	All
	3	Y	All
	4	Y	Builders/owners and certain operating company employees (superusers)
	5	N	
5.1.2 General public and bystanders*: In this case, they may be directly impacted*	1	Y	People will likely seek to be assured that system performance is at least as good as for a human alone
	2	N	
	3	N	
	4	N	
5.2.1 Validation and certification agencies and auditors	1	Y	
	2	Y	
	3	Y	
	4	Y	Assessors may wish to test performance "what if""
	5	N	
5.2.2 Incident investigators	1	Y	
	2	Y	
	3	Y	
	4	N	
	5	N	
5.2.3 Expert advisors in administrative actions or litigation	1	Y	
	2	Y	
	3	Y	
	4	N	
NOTE—While the human driver remains the final arbiter, the focus is likely to be on their professionalism and what alerts the system gives to support their decisions rather than the detail of why the system gave that alert. Detailed assessment of system performance is likely to be confined to safety/engineering professionals maintaining or developing the system. Human factors will play an important part in terms of the degree to which the driver becomes dependent on the system and potentially loses concentration.			

B.7.2.2 AI replaces the driver

In this case (see Table B.10), AI completely replaces the human driver and makes braking decisions, sounds the horn, and reports incidents. Recording and analysis may be required even where no brake or horn demand is generated to allow undetected incidents to be analyzed. Thus, the recording demands will be very high.

Table B.10—AI replacement system transparency scoresheet

System: ATO with AI oversight **Assessor: A safety engineer** **Date: xx.xx.xx**			
IEEE Std 7001-2021 subclause	**Level**	**Yes/No**	**Notes**
5.1.1 Users: *Train operating company (safety duty holder)(mix of domain expert and superusers), train owners, train builders (superusers).*	1	Y	All
	2	Y	All
	3	Y	All
	4	Y	Builders/owners and certain operating company employees (superusers)
	5	Y	Train builders/System designers
5.1.2 General public and bystanders*: In this case, they may be directly impacted*	1	Y	People will likely seek to be assured that system performance is not degraded
	2	Y	Technical press/media may demand this level of explanation
	3	N	
	4	N	
5.2.1 Validation and certification agencies and auditors	1	Y	
	2	Y	
	3	Y	
	4	Y	
	5	Y	A quantitative assessment of capability may be required
5.2.2 Incident investigators	1	Y	
	2	Y	
	3	Y	
	4	Y	
	5	Y	Particularly for undetected incidents, there will need to be an understanding of what would have changed the outcome.
5.2.3 Expert advisors in administrative actions or litigation	1	Y	
	2	Y	
	3	Y	
	4	Y	Particularly for undetected incidents, there will need to be an understanding of what would have changed the outcome.
NOTE—Striking a balance between achieving something better than a human driver and demanding similar levels of quantitative assurance to the Interlocking and ATP systems is likely to be challenging and early incidents have a high probability of being tested in court.			

B.8 Vehicle emissions measurement and mitigation system

This fictional scenario (see Table B.11) describes a case where an auto manufacturer is developing a cheaper but cleaner engine that will be capable of using either diesel or gasoline when its electric engine is depleted. The vehicle emissions subsystem is classified as Level 4, Fully Autonomous [4.2, item d)]. While vehicles involved are not driverless in today's implementation, drivers have no direct control over the functioning of this subsystem. Using this standard as a guide, they draft the following STS for the vehicle's prospective emissions measurement and mitigation system suppliers. The specification would be included in its Call for tenders/Request for proposals.

Table B.11—Vehicle emissions measurement and mitigation system transparency scoresheet

System: Vehicle Emissions Measurement and Mitigation System **Specifier:** ***Vehicle Engine Manufacturer*** **Date:** ***18 January 2020*** **Notes on overall transparency priorities:** ***Transparency is helpful for public health and well-being and for enterprises to avoid costly litigation or personal criminal liability.***					
IEEE Std 7001-2021 subclause (C = cumulative, NC = non cumulative)	**Levels (tick to indicate levels required)**				
	1	**2**	**3**	**4**	**5**
5.1.1 Users (NC)	X *	X *	X **	X **	
Users are defined as: *Drivers and classed as domain expert users. Impact on driver-operator versus driver-owner is similar but not identical. ISO 9001:2015 [B27] is not required, but a certification that the vehicle is compliant with air quality regulations and does not exceed the sustainability goals generally accepted for this class of vehicle. Any additional maintenance required (e.g., diesel exhaust fluid, filter replacement) shall be explained at Level 2 or better. **Another potential category of users, in this case, may be internal expert quality assurance and testers. These users require access to Levels 3 and 4 of transparency, even if users, i.e., drivers, do not.					
5.1.2 General public and bystanders (NC)	X	X			
NOTE—Polluted air impacts even non-driver, non-owners; this includes health impacts on children, flora, and fauna, as well as indirect, out-of-area impacts due to climate change. A public statement of the Vehicle Emissions Measurement and Mitigation System environmental impact and how it was achieved in lay terms is to be provided.					
5.2.1 Validation and certification agencies (C)	X	X	X	X	
NOTE—Evidence of validation by air quality and safety regulators is to be provided. Given that the Vehicle Emissions Measurement and Mitigation System measures as well as implements air quality and fuel efficiency controls, the supplier should deliver transparent explanations of how measurements can be externally validated against third party tools. These are to be kept current as new versions of the Vehicle Emissions Measurement and Mitigation System are released by the supplier.					
5.2.2 Incident investigators (C)	X	X	X	X	
NOTE—Resources, such as on-vehicle "black boxes," should provide sufficient data to assess Vehicle Emissions Measurement and Mitigation System compliance with its performance claims.					
5.2.3 Expert advisors in administrative actions or litigation (NC)	X	X	X	X	X
NOTE—Vehicle Emissions Measurement and Mitigation System transparency should take into account its impact on both enterprise legal counsel and external counsel in the case of litigation. This category extends to expert witnesses that may be need to provide testimonies related to the quality and testing procedures of the system.					

Annex C

(informative)

Bibliography

Bibliographical references are resources that provide additional or helpful material but do not need to be understood or used to implement this standard. Reference to these resources is made for informational use only.

[B1] Aler Tubella, A., A. Theodorou, F. Dignum, and V. Dignum, "Governance by Glass-box: Implementing Transparent Moral Bounds for AI Behaviour, Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI-2019), pp. 5787–5793.[8]

[B2] Ananny, M. and K. Crawford, "Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability," New Media & Society, vol. 20, no. 3, pp. 973–989, March 2018,[9]

[B3] Arkin, R. C., ed. Intelligent Robotics and Autonomous Agents Series. Cambridge, MA: MIT Press.[10]

[B4] Beer, J. M., A. D. Fisk, and W. A. Rogers, "Toward a framework for levels of robot autonomy in human-robot interaction," Journal of Human-Robot Interaction, vol. 3, no. 2, pp. 74–99, July 2014.[11]

[B5] Billings, C. E., "Human-Centered Aviation Automation: Principles and Guidelines," NASA Technical Memorandum 110381, Feb. 1996.[12]

[B6] Booth, S., C. Muise, and J. Shah, "Evaluating the Interpretability of the Knowledge Compilation Map: Communicating Logical Statements Effectively," Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI-2019), pp. 5801–5807.

[B7] Booth, S., Y. Zhou, A. Shah, and J. Shah, "Bayes-TrEx: a Bayesian Sampling Approach to Model Transparency by Example," Dec. 16, 2020.[13]

[B8] Bryson, J. J. and A. Theodorou, "How Society Can Maintain Human-Centric Artificial Intelligence," in Human-Centered Digitalization and Services, Toivonen, M. and E. Saari, eds. Singapore: Springer, 2019, pp. 305–323.

[B9] BS 8611:2016, Robots and robotic devices: Guide to the ethical design and application of robots and robotic systems.[14]

[B10] Bygrave, L. A., "Automated profiling: Minding the machine: Article 15 of the EC data protection directive and automated profiling," Computer Law & Security Review, vol. 17, no. 1, pp. 17–24, January 2001.[15]

[8]Available at: https://doi.org/10.24963/ijcai.2019/802
[9]Available at: https://dx.doi.org/10.1177/1461444816676645
[10]Available at: https://mitpress.mit.edu/books/series/intelligent-robotics-and-autonomous-agents-series
[11]Available at: https://doi.org/10.5898/JHRI.3.2.Beer
[12]Available at: https://ntrs.nasa.gov/api/citations/19960016374/downloads/19960016374.pdf
[13]Available at: https://arxiv.org/pdf/2002.10248.pdf
[14]Available at: https://shop.bsigroup.com/products/robots-and-robotic-devices-guide-to-the-ethical-design-and-application-of-robots-and-robotic-systems/standard
[15]Available at: https://doi.org/10.1016/S0267-3649(01)00104-2

[B11] Cappelli, C., H. Cunha, B. Gonzalez-Baixauli, and J. C. S. do Prado Leite, "Transparency versus security: early analysis of antagonistic requirements," *Proceedings of the* 2010 *ACM symposium on applied computing*, pp. 298–305.[16]

[B12] Cappelli, C., P. Engiel, R. M. Araujo, and J. C. S. do Prado Leite, "Managing Transparency Guided by a Maturity Model," 3rd Global Conference on Transparency Research HEC PARIS. October 2013.[17]

[B13] De Graaf, M. M. A. and B. F. Malle, "How People Explain Action (and Autonomous Intelligent Systems Should Too)," AAAI Fall Symposium Series 2017 AAAI Fall Symposium Series, 2017.[18]

[B14] Doshi-Velez, F. and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," Mar. 2, 2017.[19]

[B15] Dragan, A., and S. Srinivasa. "Generating Legible Motion," presented at the International conference on robotics: Science and systems, Berlin, Germany. 2013.[20]

[B16] Durst, P.J. and M. Gray. "Levels of autonomy and autonomous system performance assessment for intelligent unmanned systems," ERDC/GSL SR-14-1, 2014.[21]

[B17] Edwards, L and M. Veale, "Slave to the Algorithm: Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For," Duke Law & Technology Review, vol. 16, Dec. 2017.[22]

[B18] Endsley, M. R. and D. B. Kaber, "Level of automation effects on performance, situation awareness and workload in a dynamic control task," Ergonomics, vol. 42, no. 3, pp. 462–492, March 1999.[23]

[B19] Gilpin, L. H., D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining Explanations: An Overview of Interpretability of Machine Learning," Proceedings of the 2018 *IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pp. 80–89.

[B20] Gregor, S. and I. Benbasat, "Explanations from Intelligent Systems: Theoretical Foundations and Implications for Practice," Management Information Systems Quarterly, vol. 23, no. 4, pp. 497–530, December 1999.[24]

[B21] IEEE Ethically Aligned Design, First Edition, 2019.

[B22] IEEE Std 1012, IEEE Standard for System, Software, and Hardware Verification and Validation.

[B23] IEEE Std 1616-2021, IEEE Standard for Motor Vehicle Event Data Recorder (MVEDR).

[B24] IEEE Std 1872-2015, IEEE Standard Ontologies for Robotics and Automation.

[B25] IEEE Std 7000-2021, IEEE Standard Model Process for Addressing Ethical Concerns During System Design.

[B26] IEEE/ISO/IEC Std 15288:2015, Systems and software engineering—System life cycle processes.

[16]Available at: https://doi.org/10.1145/1774088.1774151
[17]Available at: https://www.researchgate.net/publication/282659712_Managing_Transparency_Guided_by_a_Maturity_Model
[18]Available at: https://aaai.org/ocs/index.php/FSS/FSS17/paper/view/16009/15283
[19]Available at: https://arxiv.org/pdf/1702.08608.pdf
[20]Available at: https://personalrobotics.cs.washington.edu/publications/dragan2013legible.pdf
[21]Available at: https://erdc-library.erdc.dren.mil/jspui/bitstream/11681/3284/1/ERDC-GSL-SR-14-1.pdf
[22]Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2972855
[23]Available at: https://doi.org/10.1080/001401399185595
[24]Available at: https://www.jstor.org/stable/249487

[B27] ISO 9001:2015, Quality management systems—Requirements.[25]

[B28] ISO/IEC 33000, Process Assessment.

[B29] ISO/IEC 38505-1:2017, Information technology—Governance Of IT—Governance Of Data—Part 1: Application of ISO/IEC 38500 to the Governance of Data.

[B30] ISO/IEC/IEEE 12207:2017, Systems and software engineering—Software life cycle processes.

[B31] ISO/IEC/IEEE 29119, Software and systems engineering—Software testing.

[B32] ISO/IEC/IEEE 42010, Systems and software engineering—Architecture description.

[B33] ISO/IEC/IEEE 42020, Software, systems and enterprise—Architecture processes.

[B34] Iyer, R., Y. Li, H. Li, M. Lewis, R. Sundar, and K. Sycara, "Transparency and Explanation in Deep Reinforcement Learning Neural Networks," Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 144–150, 2018.[26]

[B35] Kulesza, T. and S. Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. "Too much, too little, or just right? Ways explanations impact end users' mental models" Proceedings of IEEE Symposium on Visual Languages and Human-Centric Computing, 2013.

[B36] Lipton, Z. C., "The mythos of model interpretability," ACM Queue; Tomorrow's Computing Today, vol. 16, no. 3, pp. 31–57, May/June 2018.[27]

[B37] Mercier, S. and C. Tessier, "Some basic concepts for shared autonomy: A first report," Frontiers in Artificial Intelligence and Applications, vol. 176, pp. 40–48, 2008.[28]

[B38] NIST SP 1011-II-1.0, 2007, Autonomy Levels for Unmanned Systems (ALFUS) Framework, Volume II: Framework Models Version 1.0.[29]

[B39] Norman, D. A., "The 'problem' with automation: Inappropriate feedback and interaction, not 'over-automation,'," Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, vol. 327, no. 1241, pp. 585–593, April 1990.[30]

[B40] Olhede, S. and P. Wolfe, "When algorithms go wrong, who is liable?" Significance, vol. 14, no. 6, pp. 8–9, December 2017.[31]

[B41] Olszewska, J. I., "Designing Transparent and Autonomous Intelligent Vision Systems," Proceedings of the International Conference on Agents and Artificial Intelligence, pp. 850–856.[32]

[B42] Perlmutter, L., E. Kernfeld, and M. Cakmak, "Situated Language Understanding with Human-like and Visualization-Based Transparency," Robotics Science and Systems: Online Proceedings, 2016.[33]

[25]ISO/IEC publications are available from the ISO Central Secretariat (https://www.iso.org/). ISO/IEC publications are available in the United States from the American National Standards Institute (https://www.ansi.org/).
[26]Available at: https://doi.org/10.1145/3278721.3278776
[27]Available at: https://arxiv.org/abs/1606.03490
[28]Available at: https://ebooks.iospress.nl/publication/4217
[29]Available at: https://doi.org/10.6028/NIST.sp.1011-II-1.0
[30]Available at: https://www.jstor.org/stable/55330
[31]Available at: https://rss.onlinelibrary.wiley.com/doi/10.1111/j.1740-9713.2017.01085.x
[32]Available at: https://www.scitepress.org/Papers/2019/75852/75852.pdf
[33]Available at: http://www.roboticsproceedings.org/rss12/p40.pdf

[B43] Rose, K. H., Project Quality Management: Why, What and How. Fort Lauderdale, FL: J. Ross Publishing, 2014.

[B44] Rotsidis, A., A. Theodorou, J. J. Bryson, and R. H. Wortham, "Improving Robot Transparency: An Investigation With Mobile Augmented Reality," 2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), pp. 1–8.

[B45] SAE J3016_201806, Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles.[34]

[B46] Sampaio do Prado Leite, J. C. and C. Cappelli, "Software Transparency," Business & Information Systems Engineering, vol. 2, no. 3, pp. 127–139, 2010.[35]

[B47] Scherer, M. U., "Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies," Harvard Journal of Law & Technology, vol. 29, no. 2, Spring 2015.[36]

[B48] Schifers, C. and G. Hans, "IEC 61375-1 and UIC 556—International standards for train communication," 2000 IEEE Conference on Vehicular Technology (VTC), pp. 1581–1585.

[B49] Selbst, A. D., A Mild Defense of Our New Machine Overlords, Vol. 70. Vanderbilt Law Review, 2017.[37]

[B50] Sheridan, T. B., Telerobotics, Automation, and Human Supervisory Control. Cambridge, MA: MIT Press, 1992.

[B51] Sheridan, T. B. and W. L. Verplank, "Human and computer control of undersea teleoperators, Technical Report 15, Massachusetts Institute of Technology Man-Machine Systems Lab, Cambridge, MA, 1978.

[B52] Shneiderman, B., "Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy," International Journal of Human-Computer Interaction, vol. 36, no. 6, pp. 495–504, March 2020.[38]

[B53] Skraaning, G. and G. A. Jamieson, "Human Performance Benefits of The Automation Transparency Design Principle: Validation and Variation," Human Factors, pp. 1–23, 2019.[39]

[B54] Stepney, S. and F. A. C. Polack, Engineering Simulations as Scientific Instruments: A Pattern Language. Cambridge: Springer International Publishing, 2018.[40]

[B55] Theodorou, A., "AI governance through a transparency lens," PhD dissertation, University of Bath, 2019.

[B56] Theodorou, A., R. H. Wortham, and J. J. Bryson, "Designing and implementing transparency for real time inspection of autonomous robots," Connection Science, vol. 29, no. 3, pp. 230–241, 2017.[41]

[B57] Wachter, S., B. Mittelstadt, and C. Russell, "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR," Harvard Journal of Law & Technology, vol. 31, no. 2, October 2017.[42]

[34]SAE publications are available from the Society of Automotive Engineers (http://www.sae.org/).
[35]Available at: https://link.springer.com/article/10.1007%2Fs12599-010-0102-z
[36]Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2609777
[37]Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2941078
[38]Available at: https://www.tandfonline.com/doi/full/10.1080/10447318.2020.1741118
[39]Available at: https://journals.sagepub.com/doi/10.1177/0018720819887252
[40]Available at: https://link.springer.com/book/10.1007/978-3-030-01938-9
[41]Available at: https://www.tandfonline.com/doi/full/10.1080/09540091.2017.1310182
[42]Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3063289

[B58] Walsh, T., "Turing's red flag," Communications of the ACM, vol. 59, no. 7, pp. 34–37, July 2016.[43]

[B59] Winfield, A. F. T. and M. Jirotka, "The Case for an Ethical Black Box," in Towards Autonomous Robotic Systems, Gao, Y., S. Fallah, Y. Jin, and C. Lekakou, eds. Cambridge: Springer, 2017, pp. 262–273.[44]

[B60] Winfield, A. F. T. and M. Jirotka, "Ethical governance is essential to building trust in robotics and artificial intelligence systems," Philosophical Transactions of the Royal Society A, 2018.[45]

[B61] Winfield, A. F. T., K. Winkle, H. Webb, U. Lyngs, M. Jirotka, and C. Macrae, "Robot Accident Investigation: a case study in Responsible Robotics," in Software Engineering for Robotics, Cavalcanti, A., B. Dongol, R. Hierons, J. Timmis, and J. Woodcock, eds. Cambridge: Springer, 2021.[46]

[B62] Wortham, R.H., Transparency for Robots and Autonomous Systems: Fundamentals, technologies and applications. London: The Institution of Engineering and Technology, 2020.[47]

[B63] Wortham, R. H., A. Theodorou, and J. J. Bryson, "Improving robot transparency: Real-time visualisation of robot AI substantially improves understanding in naive observers," 2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), pp. 1424–1431.[48]

[B64] Zarsky, T., "Transparent Predictions," University of Illinois Law Review, no. 4, pp. 1503–1570, 2013.[49]

[43] Available at: https://dl.acm.org/doi/10.1145/2838729
[44] Available at: https://link.springer.com/chapter/10.1007%2F978-3-319-64107-2_21
[45] Available at: https://royalsocietypublishing.org/doi/10.1098/rsta.2018.0085
[46] Available at: https://link.springer.com/chapter/10.1007/978-3-030-66494-7_6
[47] Available at: https://digital-library.theiet.org/content/books/ce/pbce130e
[48] Available at: https://ieeexplore.ieee.org/document/8172491
[49] Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2324240

IEEE SA
STANDARDS ASSOCIATION

RAISING THE WORLD'S STANDARDS

Connect with us on:

Twitter: twitter.com/ieeesa
Facebook: facebook.com/ieeesa
LinkedIn: linkedin.com/groups/1791118
Beyond Standards blog: beyondstandards.ieee.org
YouTube: youtube.com/ieeesa

standards.ieee.org
Phone: +1 732 981 0060

IEEE