

Ethical Risk Assessment and Transparency in 2 Standards

BS8611 and IEEE 7001

Alan FT Winfield
Bristol Robotics Laboratory
alan.winfield@brl.ac.uk

Talk for Dependable and Deployable AI for Robotics
University of Edinburgh
26 November 2025

BS 8611:2016



BSI Standards Publication

Robots and robotic devices

Guide to the ethical design and
application of robots and robotic
systems

Ethical Risk Assessment

BS8611

- BS8611 sets out 20 distinct *ethical hazards and risks*, grouped under four categories:
 - societal
 - application
 - commercial/financial
 - environmental
- Advice on measures to mitigate the impact of each risk is given, along with suggestions on how such measures might be verified or validated

Ethical Harms, Hazards and Risks

- An ethical **harm** is defined as
 - that which might compromise psychological and/or societal and environmental well-being
- An ethical **hazard** is defined as
 - a possible source of ethical harm
- An ethical **risk** is defined as
 - a possible ethical harm arising from exposure to a hazard

Some societal hazards risks & mitigation from BS8611

Hazard	Risk	Mitigation
Deception	Confusion; unintended (perhaps) delayed consequences, such as loss of trust.	Design robot behaviour & appearance to avoid deception. Ensure transparency of robotic nature.
Anthropomorphisation.	Misinterpretation.	Avoid unnecessary human-like design. (See also deception)
Privacy & confidentiality	Unauthorized access, collection and/or distribution of personal data.	Data control. Ensure clear justification for data collection, user awareness and informed consent.
Addiction	Loss of human capability, dependency on robot, reduced willingness to engage with others, isolation	Raise awareness of over-dependency.

RoboTed: a worked example

- Consider a fictional robot toy called RoboTed
 - RoboTed is an Internet (WiFi) connected device,
 - RoboTed has cloud-based speech recognition and conversational AI (chatbot) and local speech synthesis,
 - RoboTed's eyes are functional cameras allowing RoboTed to recognise faces,
 - RoboTed has motorised arms and legs to provide it with limited baby-like movement and locomotion.



What can RoboTed do?

- RoboTed is designed to:
 - *Recognise* its owner, learning their face and name and turning its face toward the child.
 - *Respond* to physical play such as hugs and tickles.
 - *Tell stories* while allowing a child to interrupt the story to ask questions or ask for sections to be repeated.
 - *Sing songs*, while encouraging the child to sing along and learn the song.
 - Act as a *child minder*, allowing parents to both remotely listen, watch and speak via RoboTed.

Ethical Risk Assessment 1 psychological risks

Hazard	Risk		Mitigation
Addiction	Child plays with RoboTed obsessively and neglects family	M	Explore 'RoboTed needs to sleep' function
Deception (of child)	Child believes that RoboTed has feelings (for her)	M	Design chatbot to avoid language that suggests feelings
Over trusting (by parents)	Parents come to rely on the child minder function	H	Remove the child minder function
 The Uncanny Valley	Child becomes fearful of robot	L	Use 'cartoon' voice; engage children in early user trials

Ethical Risk Assessment 2 privacy & transparency risks

Hazard	Risk		Mitigation
Weak security	Malicious hackers gain access to RoboTed's sensors & control functions.	H	Implement strong encryption together with best practice password protection.
Privacy	Personal data, including images and voice recordings of child are stolen.	M	Put in place auditable measures to ensure personal data is deleted immediately
Lack of transparency	Lack of data logs makes it hard or impossible to investigate accidents	H	Build a secure local data logger into RoboTed.

RoboTed ethical risk assessment: summary

- Attention to ethical risks can
 - suggest new functions, such as “RoboTed needs to sleep now”,
 - draw attention to how designs can be modified to mitigate some risks,
 - highlight the need for user engagement, and
 - even result in product functions being rejected as too risky.

Find Standards

Develop Standards

Get Involved

News & Events

About Us

Buy Standards

eTools

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems

An incubation space for new standards and solutions, certifications and consensus building for ethical implementation of intelligent technologies

advanced for the Benefit of
Humanity



INDUSTRY CONNECTIONS

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems

Videos & Webinars

News & Events

Ethically Aligned Design,
Version 1, Translations and
Reports

VIEW THE COMPLETE LIST

ABOUT

To ensure every stakeholder involved in the design and development of autonomous and intelligent systems is educated, trained, and empowered to prioritize ethical considerations so that these technologies are advanced for the benefit of humanity.

- [View specifics regarding the Mission and deliverables for the Initiative.](#)
- [See a list of The Initiative's Executive and other Committees.](#)
- [Learn more from Frequently Asked Questions.](#)

ETHICS IN ACTION

We've launched the second version of *Ethically Aligned Design!* [View Launch Details.](#)

Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (EA/ID) represents the collective work of several hundred participants from six continents who are thought

<https://ethicsinaction.ieee.org/>

Box 1 | IEEE P7000 series human standards in development

- P7000 — *Model Process for Addressing Ethical Considerations*
- P7001 — *Transparency of Autonomous Systems*
- P7002 — *Data Privacy Process*
- P7003 — *Algorithmic Bias Considerations*
- P7004 — *Standard on Child and Student Data Governance*
- P7005 — *Standard on Employer Data Governance*
- P7006 — *Standard on Personal Data Artificial Intelligence*
- P7007 — *Ontological Standard for Ethically Driven Systems*
- P7008 — *Standard for Ethically Driven Nudging Systems*
- P7009 — *Standard for Fail-Safe Design of Autonomous Systems*
- P7010 — *Wellbeing Metrics Standard for Ethical Systems*
- P7011 — *Standard for the Process of Identifying and Managing News Sources*
- P7012 — *Standard for Machine Readable Personhood*
- P7013 — *Inclusion and Application Standards for Artificial Intelligence*
- P7014 -- *Std for Ethical considerations in Emulated Systems*
- P7015 -- *Standard for Data and Artificial Intelligence*



ETHICALLY ALIGNED DESIGN

First Edition Overview

A Vision for Prioritizing Human Well-being
with Autonomous and Intelligent Systems



Box 1 | IEEE P7000 series human standards in development

P7000 — *Model Process for Addressing Ethical Concerns During System Design*

P7001 — *Transparency of Autonomous Systems*

P7002 — *Data Privacy Process*

P7003 — *Algorithmic Bias Considerations*

P7004 — *Standard on Child and Student Data Governance*

P7005 — *Standard on Employer Data Governance*

P7006 — *Standard on Personal Data Artificial Intelligence (AI) Agent*

P7007 — *Ontological Standard for Ethically Driven Robotics and Automation Systems*

P7008 — *Standard for Ethically Driven Nudging for Robotic, Intelligent and Autonomous Systems*

P7009 — *Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems*

P7010 — *Wellbeing Metrics Standard for Ethical Artificial Intelligence and Autonomous Systems*

P7011 — *Standard for the Process of Identifying and Rating the Trustworthiness of News Sources*

P7012 — *Standard for Machine Readable Personal Privacy Terms*

P7013 — *Inclusion and Application Standards for Automated Facial Analysis Technology*

P7014 -- *Std for Ethical considerations in Emulated Empathy in Autonomous and Int. Systems*

P7015 -- *Standard for Data and Artificial Intelligence (AI) Literacy, Skills, and Readiness*

IEEE Std 7001-2021

Winfield AFT, Booth S, Dennis LA, Egawa T, Hastie H, Jacobs N, Muttram RI, Olszewska JI, Rajabiyazdi F, Theodorou A, Underwood MA, Wortham RH and Watson E (2021)
IEEE P7001: A Proposed Standard on Transparency. *Front. Robot. AI* 8:665729.
doi: 10.3389/frobt.2021.665729

IEEE Standard for Transparency of Autonomous Systems

IEEE Vehicular Technology Society

IEEE Robotics and Automation Society

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

IEEE Std 7001™-2021



The origin of 7001

- One of the 8 General Ethical Principles set out in IEEE SA's [Ethically Aligned Design](#), is the principle of transparency:

The basis of a particular autonomous or intelligent system decision should always be discoverable.

- IEEE Working group P7001 was initiated as a direct response to this principle

The aim of 7001

- The aim of IEEE standard 7001 is to set out **measurable, testable** levels of transparency, so that autonomous systems can be objectively assessed, and levels of compliance determined.

How does 7001 define transparency?

- What do we mean by *transparency in autonomous and intelligent systems*?
- A system is *transparent* if it is *possible to discover why it behaves in a certain way*, for instance, why it made a particular decision.
 - A system is *explainable* if the way it behaves can be expressed in plain language understandable to non-experts.

How does 7001 define transparency?

Formal definition in 7001:

“the **transfer of information** from an autonomous system or its designers to a **stakeholder**, which is **honest**, **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”

non-experts.

But transparency is not one thing

- Transparency means something different to different *stakeholders*. 7001 covers five grouped into two categories:
- Expert stakeholders:
 - *Safety certification engineers or agencies*
 - *Accident investigators*
 - *Lawyers or expert witnesses*
- Non-expert stakeholders:
 - *Users*
 - *Wider society*



Transparency for Accident/Incident Investigators

- What information does an accident investigator need to find out **why** an accident happened?
 - Details of the events leading up to the accident
 - Details of the internal decision-making process in the intelligent system.
- Established and trusted processes of air accident investigation provide an excellent model of good practice.
 - Consider the **aircraft black box** (flight data recorder).*

Example of normative requirements for Users

Transparency level Users

0 None

1 A user manual shall be provided, which sets out how a robot will behave in different circumstances

2 The manual shall be presented as an interactive visualization or simulation

3 The robot shall be equipped with a “why did you just do that?” function

4 The robot shall be equipped with a “what would you do if ...?” function

5 Not defined



7001 Processes

- System Transparency Assessment (STA)
 - a method for **assessing the transparency** of an existing system
- System Transparency Specification (STS)
 - a method for **specifying the transparency requirements** of a system prior to its development



An example STA for RoboTED

Stakeholder Group	Transparency level(s)	Evidenced by
Users	1,2	A user manual is provided for parents (level 1) An interactive online visual guide is also provided, for both parents and children (level 2)
General Public	1	P7001 level 1 requires that a robot identifies itself as an autonomous system. When powered up RoboTED announces itself as a robot
Certification agencies	2	RoboTED has been certified as safe against standard EU EN 62115 (2020) Safety of Electric Toys, and descriptions of the system and how it has been validated are available for safety certifiers. This meets 7001 level 2
Incident Investigators	2	The robot is equipped with a data logging system capable of recording sensor inputs, user commands, and actuator outputs (level 2)
Lawyers and Expert Witnesses	2	7001 level 2 requires that a system has been subjected to an ethical risk assessment, which can be made available. This is the case for RoboTED

confidence

How do we build ~~trust~~?

- We trust our technology not (just) because it is cool and convenient, but because of **Standards**, **Safety Certification** and **Regulation**
- Without transparent and robust **governance frameworks** there will be no confidence



Thank you!

- The British Standards Association and members of working group AMT/010
- The IEEE Standards Association and Members of IEEE P7001 working group
- Winfield, A.F.T., Studley, M. (2024). On the Relationship Between Benchmarking, Standards and Certification in Robotics and AI. In: Aldinhas Ferreira, M.I. (eds) Producing Artificial Intelligent Systems. Studies in Computational Intelligence, vol 1150. Springer, Cham.
https://doi.org/10.1007/978-3-031-55817-7_1 and <https://arxiv.org/pdf/2309.12139>



Some of the volunteers meet
to work on IEEE 7001