Bristol Robotics Laboratory

University of the West of England

University of BRISTOL

Ethical Standards in robotics and AI

Responsible Robotics

Alan FT Winfield
Bristol Robotics Laboratory
alanwinfield.blogspot.com
@alan_winfield

RoboSoft:
Software Engineering for Robotics
Royal Academy of Engineering
13-14 November 2019

Imagine something happens...

Bristol Robotics Laboratory

Outline

- Introduction
  - All standards embody a *principle*
  - Introducing *explicitly* ethical standards
  - From ethical principles to ethical standards
- BS8611: *the world's first explicitly ethical standard?*
- The IEEE P700X human standards in draft
  - A case study: P7001 *Transparency of Autonomous Systems*
- Responsible Robotics
  - And why we need *robot accident investigation*

Standards are infrastructure

brl
Bristol Robotics Laboratory

All standards embo

- *Safety*: the general pri d systems should
- *Quality*: the pri t practice leads to improved
- *Interoperabi* dard ways of doing things benefit all
- *All* standards embody the the values of cooperation and harmonisation

ISO 13482
Safety requirements for personal care robots

ISO 9001
Requirements for a Quality Management System

IEEE 802.11
protocols for implementing a wireless local area network

All Standards are *implicit* ethical standards

brl
Bristol Robotics Laboratory

Explicit ethical standards

- Let us … ne that addre…
- Would…
  - *thr*… pefully *rea*… otential for… es

Four categories of ethical harm:

- Unintended physical harm
- Unintended psychological harm
- Unintended socio/economic harm
- Unintended environmental harm

The Good News:
a new generation of explicitly ethical standards is now emerging

brl
Bristol Robotics Laboratory

From ethical principles to ethical standards*

Emerging Ethics:
Roboethics roadmap (2006)
EPSRC/AHRC principles (2010)
IEEE Global Initiative (2016)
plus many others…

Emerging ethical standards:
BS 8611
IEEE P700X

Emerging regulation:
Driverless cars?
Assistive robotics?
Drones?

*Winfield, A. F. and Jirotka, M. (2018) Ethical governance is essential to building trust in robotics and AI systems. Philosophical Transactions A: Mathematical, Physical and Engineering Sciences, 376 (2133). ISSN 1364-503X Available from: http://eprints.uwe.ac.uk/37556

Bristol Robotics Laboratory

A proliferation of principles

Robots and AIs should:

1. do no harm, while being free of bias and deception;
2. respect human rights and freedoms, including dignity and privacy, while promoting well-being; and
3. be transparent and dependable while ensuring that the locus of responsibility and accountability remains with their human designers or operators.

* http://alanwinfield.blogspot.com/2019/04/an-updated-round-up-of-ethical.html

BS 8611:2016

BSI Standards Publication

Ethical Risk Assessment

Robots and robotic devices

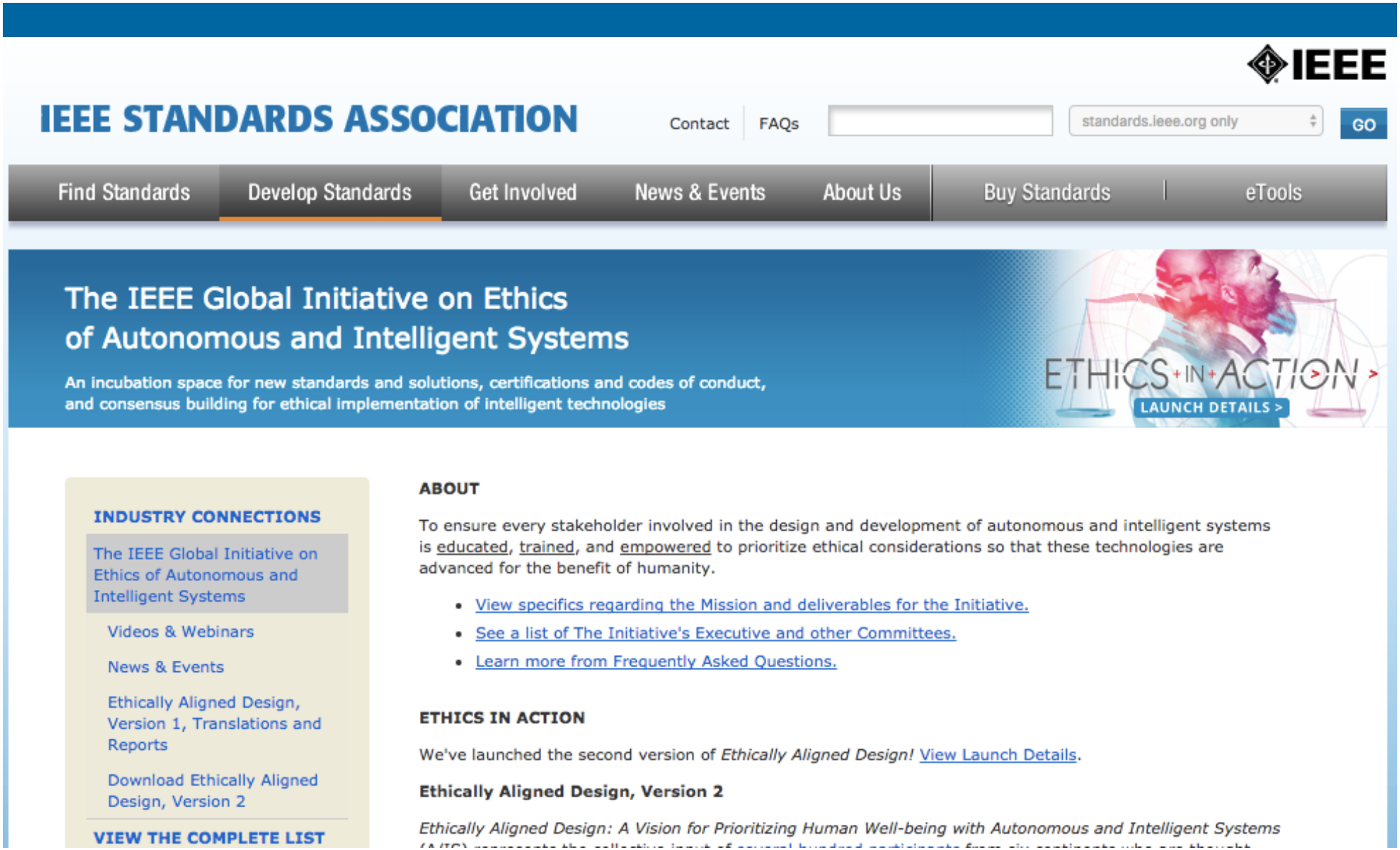
Guide to the ethical design and application of robots and robotic systems

Ethical Risk Assessment

- BS8611 articulates a set of 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

Some societal *hazards risks & mitigation*

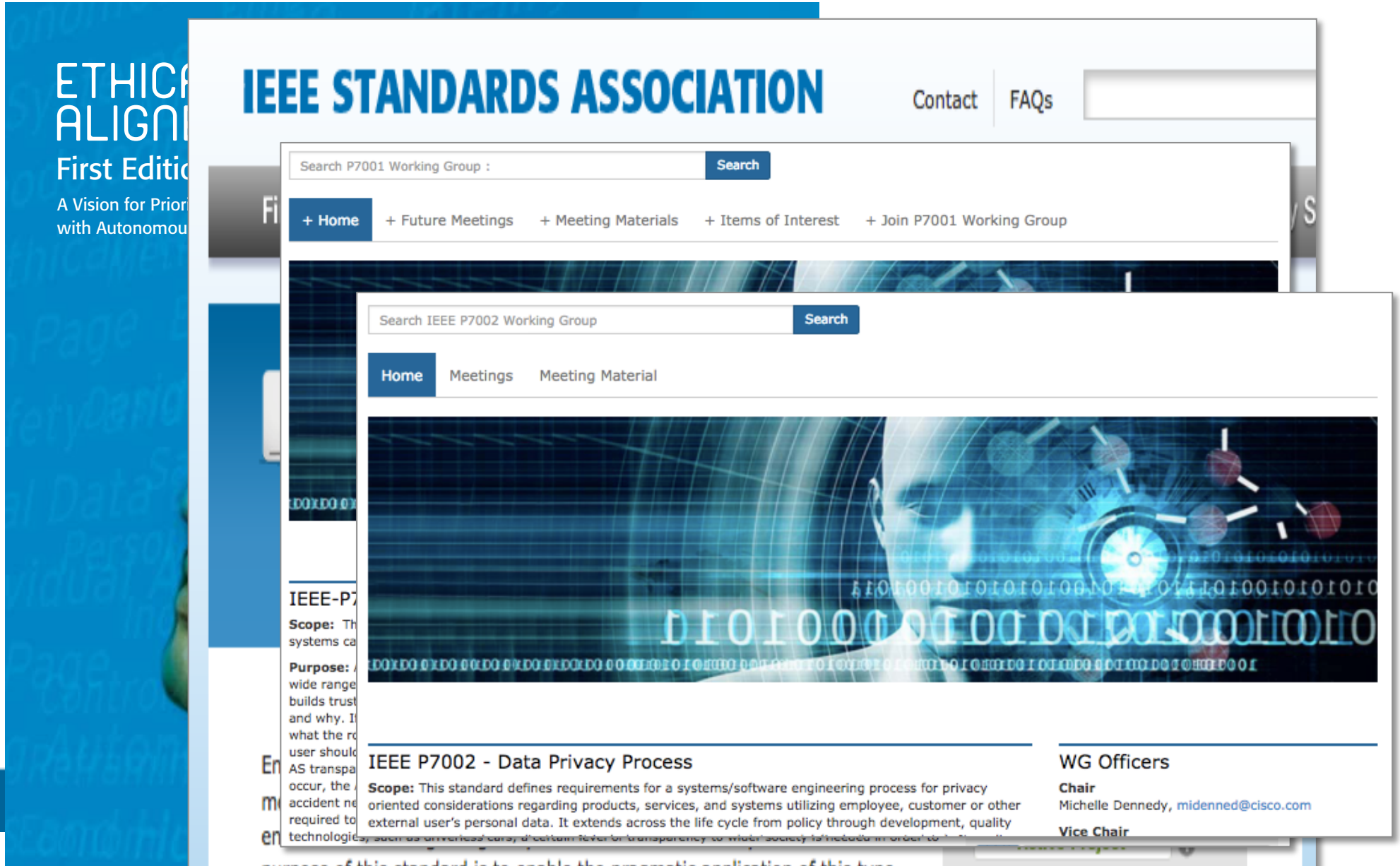
			happened	
Deception (intentional or unintentional)	Confusion, unintended (perhaps delayed) consequences, eventual loss of trust	Avoid deception due to the behaviour and/or appearance of the robot and ensure transparency of robotic nature	–	Software verification; user validation; expert guidance
Anthropo-morphization	Misinterpretation	Avoid unnecessary anthropomorphization Clarification of intent to simulate human or not, or intended or expected behaviour	See deception (above) Use anthropomorphization only for well-defined, limited and socially-accepted purposes	User validation; expert guidance
Privacy and confidentiality	Unauthorized access, collection and/or distribution of data, e.g. coming into the public domain or to unauthorized, unwarranted entities	Clarity of function Control of data, justification of data collection and distribution Ensure user awareness of data management and obtain informed consent in appropriate contexts	Privacy by design Data encryption, storage location, adherence to legislation	Software verification
Lack of respect for cultural diversity and pluralism	Loss of trust in the device, embarrassment, shame, offence	Awareness of cultural norms incorporated into programming	Organizational, professional, regional	Software verification; user validation
Robot addiction	Loss of human capability, dependency, reduction in willingness to engage with others, isolation	Raise awareness of dependency	A difficult area, particularly in relation to vulnerable people Careful evaluation of potential applications is needed	User validation; expert guidance



https://ethicsinaction.ieee.org/

Deliverables

IEEE
Advancing Technology for Humanity



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*

P7001: Transparency in autonomous systems

- What do we mean by transparency in autonomous and intelligent systems?
- A system is considered to be 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.

Why is transparency important?

- *All* robots and AIs are designed to work for, with or alongside humans – who need to be able to understand *what* they are doing and *why*
  - Without this understanding those systems will not be trusted
- Robots and AIs can and do go wrong. When they do it is *very* important that we *can find out why*.
  - Without transparency finding out what went wrong and why is extremely difficult

Transparency is not one thing

- Transparency means something different to different *stakeholders*
  - An elderly person doesn't need to understand what her care robot is doing in the same way as the engineer who repairs it
- Expert stakeholders:
  - *Safety certification engineers or agencies*
  - *Accident investigators*
  - *Lawyers or expert witnesses*
- Non-expert stakeholders:
  - *Users*
  - *Wider society*

Transparency for Accident 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 robot or AI.
- Established and trusted processes of air accident investigation provide an excellent model of good practice for autonomous and intelligent systems.
  - Consider the aircraft black box (flight data recorder).

Transparency for users

- Users need the kind of explainability that builds trust
  - By providing simple ways to understand what the system is doing, and why.
- For example:
  - The ability to ask a robot or AI why did you just do that? and receive a simple natural language explanation.
  - A higher level of user transparency would be the ability for a user to ask the system what would you do if . . . ? and receive an intelligible answer.

Transparency by Design

- How do we *design* systems to be transparent for all of the stakeholder groups above?
- We need:
  - *Process* standards for transparency, i.e. transparent and robust human processes of design, manufacture, test, deployment etc
  - *Technical* standards for transparency, i.e. requirements for *transparency*, such as P7001
  - *Technologies* for transparency, i.e. event data recorders

Responsible Innovation

- Responsible Innovation (RI) is a set of good practices for ensuring that research and innovation benefits society and the environment

For RI frameworks see
https://www.rri-tools.eu/
https://www.orbit-rri.org/ &
https://epsrc.ukri.org/research/framework/area/

The 6 pillars of RI

Responsible Robotics

The application of Responsible Innovation in the design, manufacture, operation, repair and end-of-life recycling of robots, that seeks the most benefit to society and the least harm to the environment

brl
Bristol Robotics Laboratory

RoboTIPS

robotips.co.uk

Apps | UWE email | UWE Agresso exp... | WEF TopLink | Free Calorie Coun... | Holiday | duckduckgo.com | Interpretable Mac... | Other Bookmarks

RoboTIPS

Home | Project Information | Ethical Black Box | Team | Partners

RoboTIPS

RESPONSIBLE ROBOTS FOR THE DIGITAL ECONOMY

There is often a tension between the economic needs for increasing technological innovation and the ways in which these innovations may be developed **responsibly** - that is in a manner that is societally acceptable and desirable.

In the RoboTIPS project we develop an approach that aims to anticipate not only the positive outcomes but also the potentially negative consequences of technological innovations for society. In particular we focus on the domain of **social robots.**

www.robotips.co.uk

UNIVERSITY OF OXFORD

brl Bristol Robotics Laboratory

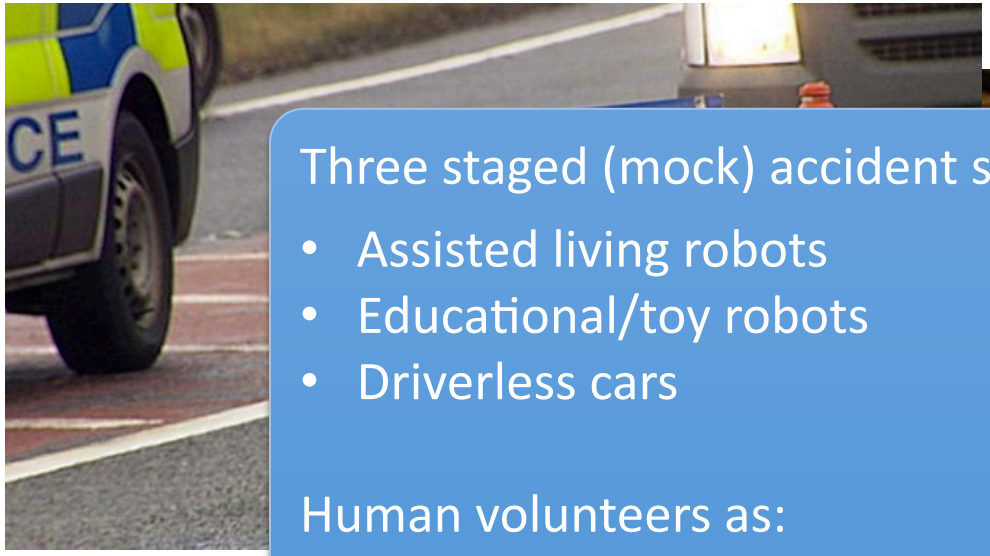
The ethical black box

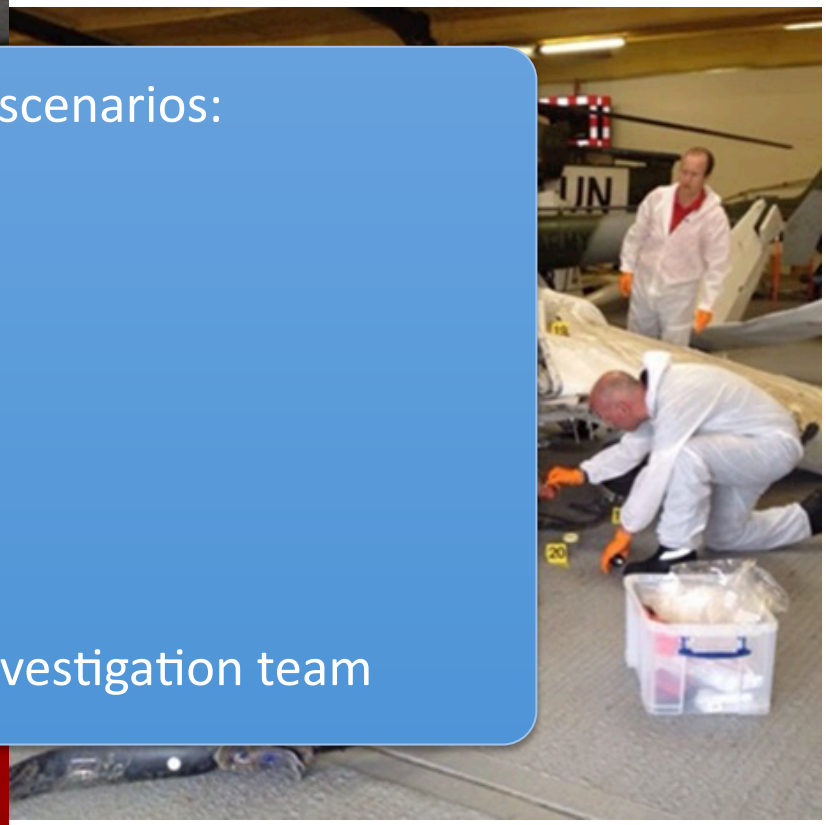
Ethical black box

AF Winfield and M Jirotka (2017) The case for an ethical black box, Towards Autonomous Robotic Systems (TAROS), LNCS 10454, 262-273

brl
Bristol Robotics Laboratory

A human process





Three staged (mock) accident scenarios:

- Assisted living robots
- Educational/toy robots
- Driverless cars

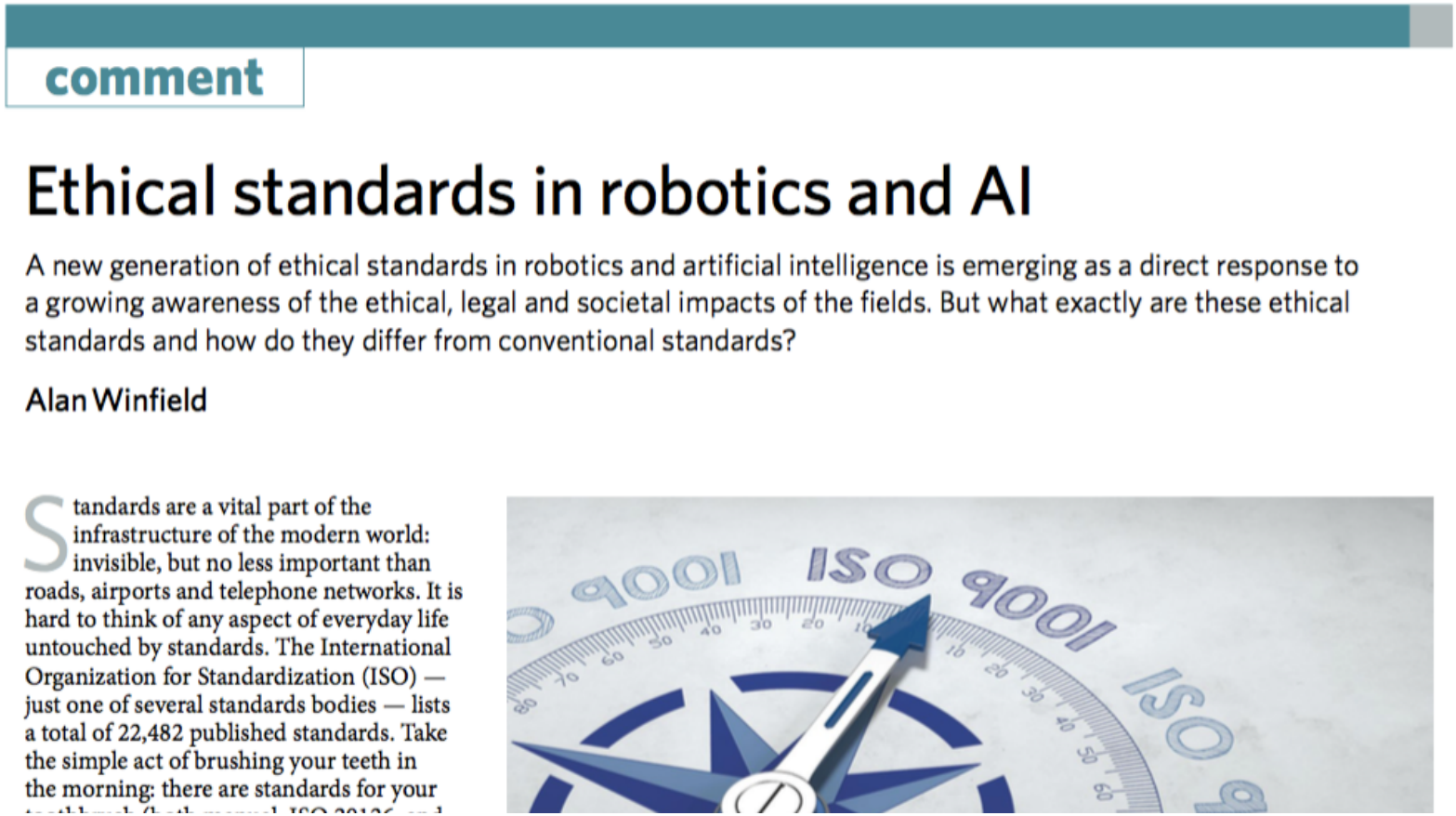
Human volunteers as:

- Subjects of the accident
- Witnesses to the accident
- Members of the accident investigation team

brl
Bristol Robotics Laboratory

Thank you!

- Ethical Standards *matter* because a new generation of social robots has *ethical* as well as safety impact



comment

Ethical standards in robotics and AI

A new generation of ethical standards in robotics and artificial intelligence is emerging as a direct response to a growing awareness of the ethical, legal and societal impacts of the fields. But what exactly are these ethical standards and how do they differ from conventional standards?

Alan Winfield

Standards are a vital part of the infrastructure of the modern world: invisible, but no less important than roads, airports and telephone networks. It is hard to think of any aspect of everyday life untouched by standards. The International Organization for Standardization (ISO) — just one of several standards bodies — lists a total of 22,482 published standards. Take the simple act of brushing your teeth in the morning: there are standards for your