Ethical Governance is needed to build Trust in robotics and AI

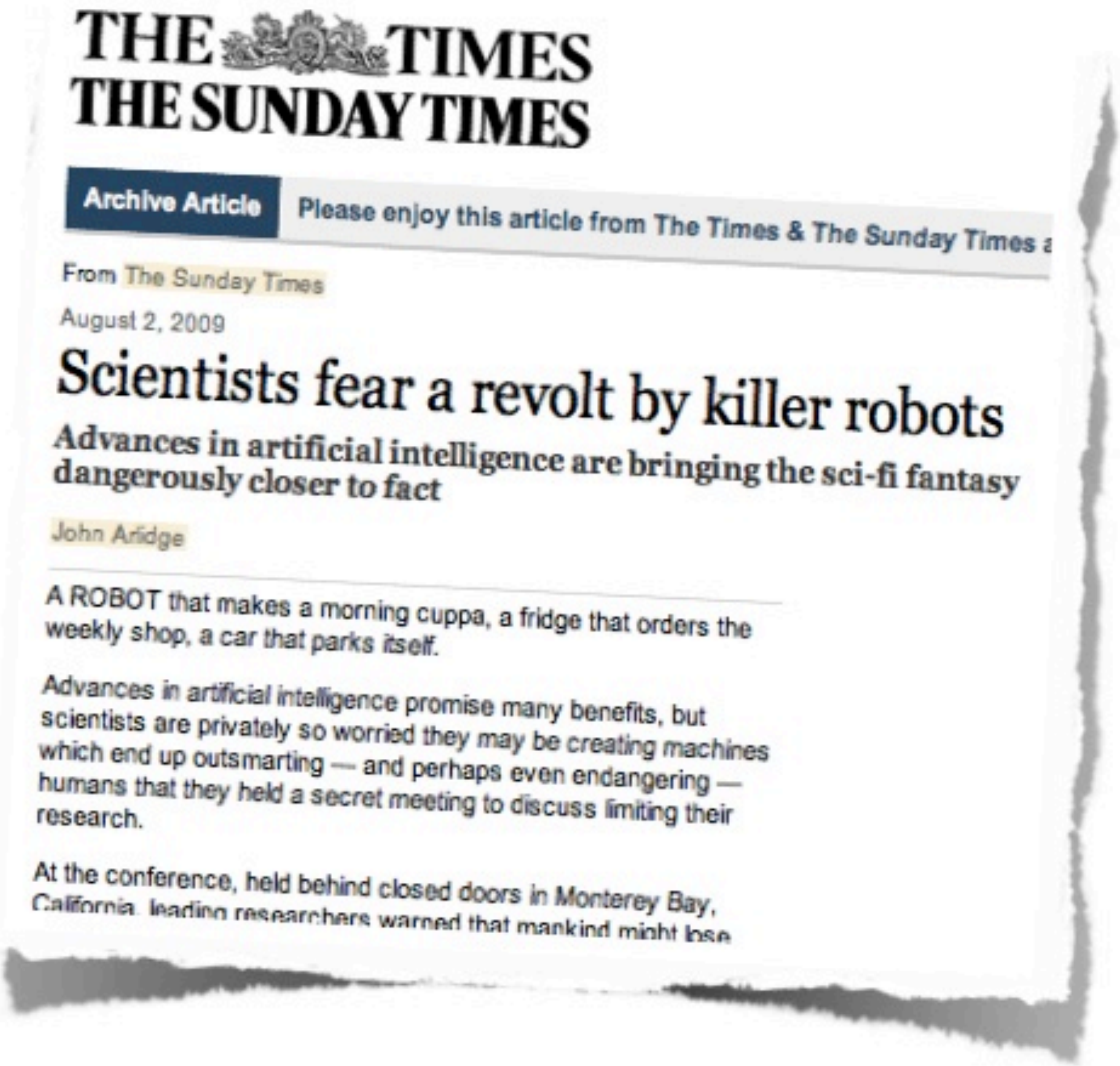
A framework for ethical governance

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

Jean Golding Institute
University of Bristol
30 January 2019

Outline

- A roadmap for ethical governance
  - From principles to regulation, and *trust*
- British Standard BS8611
  - Ethical risk assessment
- The IEEE Standards Association global ethics initiative
  - IEEE P7001 Transparency in Autonomous Systems
- From principles to practice



THE TIMES
THE SUNDAY TIMES

Archive Article Please enjoy this article from The Times & The Sunday Times a

From The Sunday Times

August 2, 2009

Scientists fear a revolt by killer robots

Advances in artificial intelligence are bringing the sci-fi fantasy dangerously closer to fact

John Arlidge

A ROBOT that makes a morning cuppa, a fridge that orders the weekly shop, a car that parks itself.

Advances in artificial intelligence promise many benefits, but scientists are privately so worried they may be creating machines which end up outsmarting — and perhaps even endangering — humans that they held a secret meeting to discuss limiting their research.

At the conference, held behind closed doors in Monterey Bay, California, leading researchers warned that mankind might lose

How do we build trust?

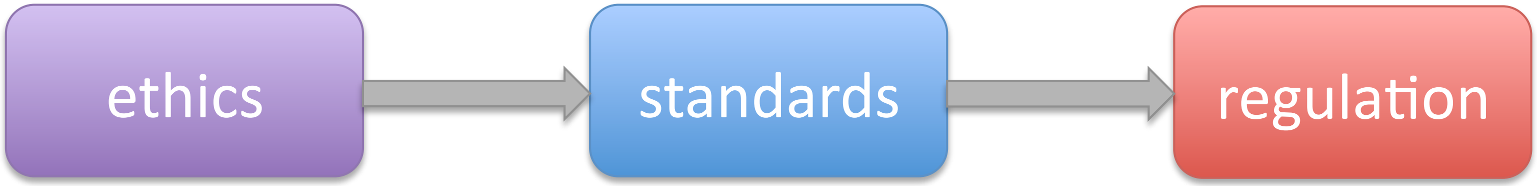
- We trust our technology not (just) because it is cool and convenient, but because of Standards, Safety Certification and Regulation
- Without transparent and robust governance frameworks there will be no trust

Build on a foundation of ethics*

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

Emerging standards:
ISO 13482
BS 8611
IEEE P700X

Emerging regulation:
Drones?
Driverless cars?
Assistive robotics?



ethics

standards

regulation

Projects:
RoboLaw

*Winfield, A. F. (2016) Written evidence submitted to the UK Parliamentary Select Committee on Science and Technology Inquiry on Robotics and Artificial Intelligence. Discussion Paper. Science and Technology Committee (Commons), Website. Available from: http://eprints.uwe.ac.uk/29428

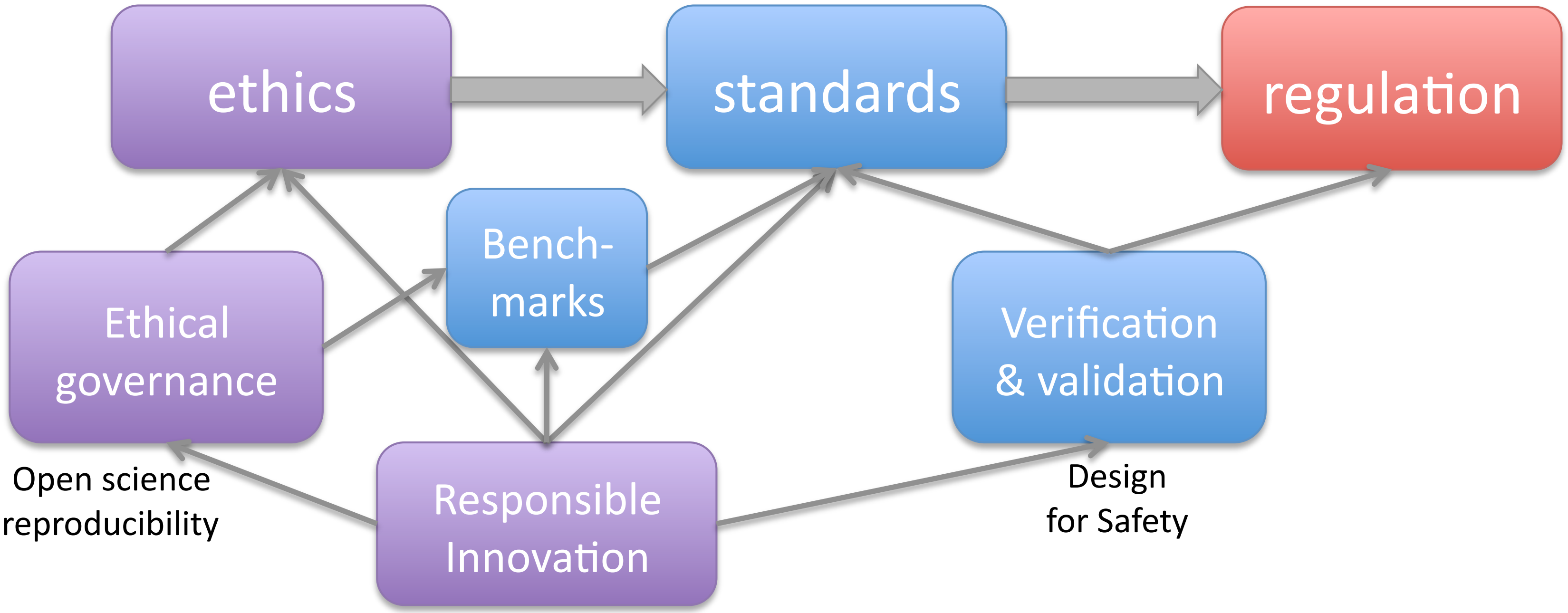
**http://alanwinfield.blogspot.co.uk/2017/12/a-round-up-of-robotics-and-ai-ethics.html

Scaffolded by
Responsible Research and Innovation

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

Emerging standards:
ISO 13482
BS 8611
IEEE P700X

Emerging regulation:
Drones?
Driverless cars?
Assistive robotics?

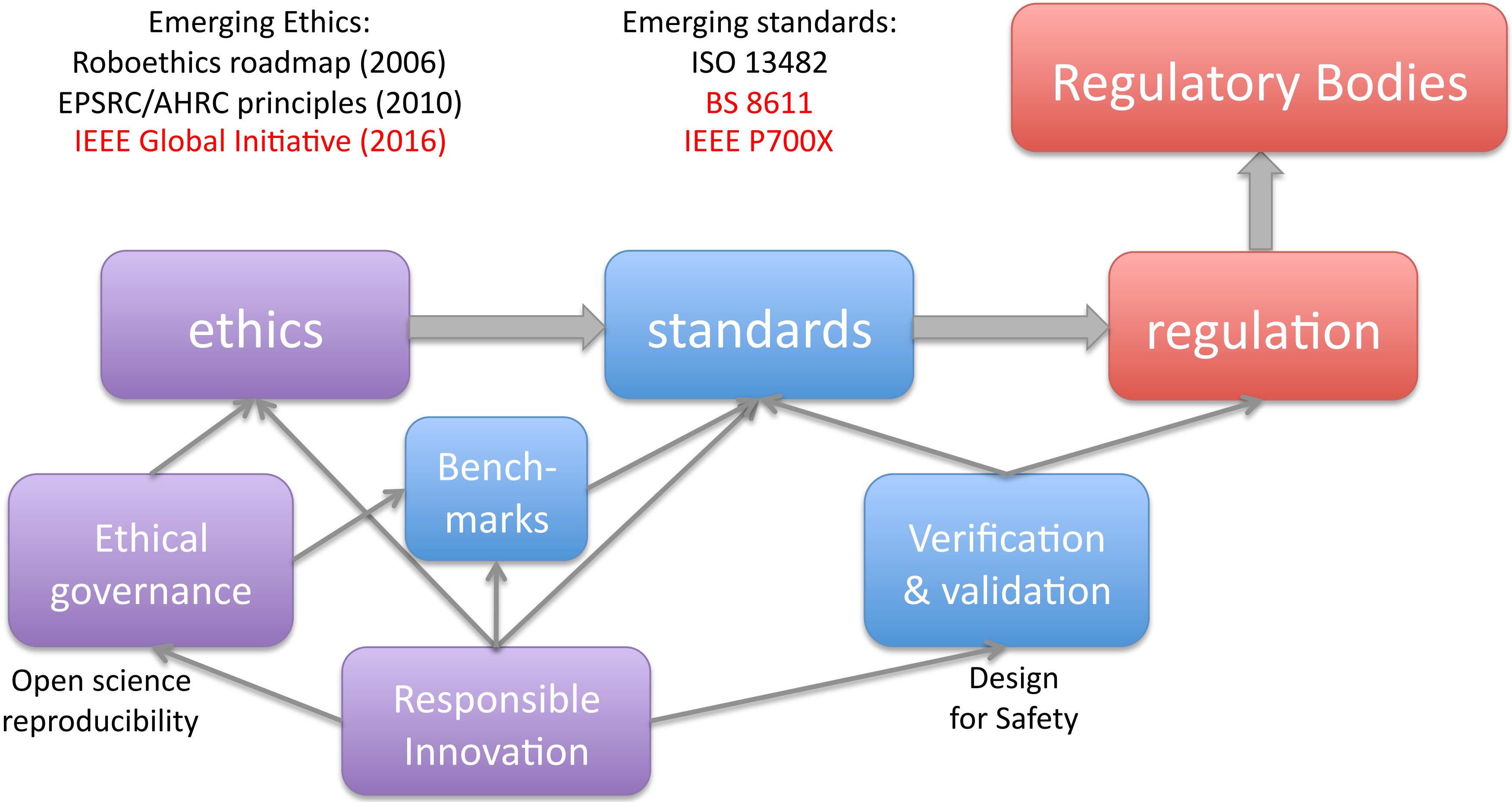


Open science
reproducibility

Design
for Safety

Rome declaration on RRI
EPSRC AREA framework
Foundation for Responsible Robotics

Regulation needs teeth



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

Emerging standards:
ISO 13482
BS 8611
IEEE P700X

Regulatory Bodies

ethics

standards

regulation

Bench-marks

Ethical governance

Verification & validation

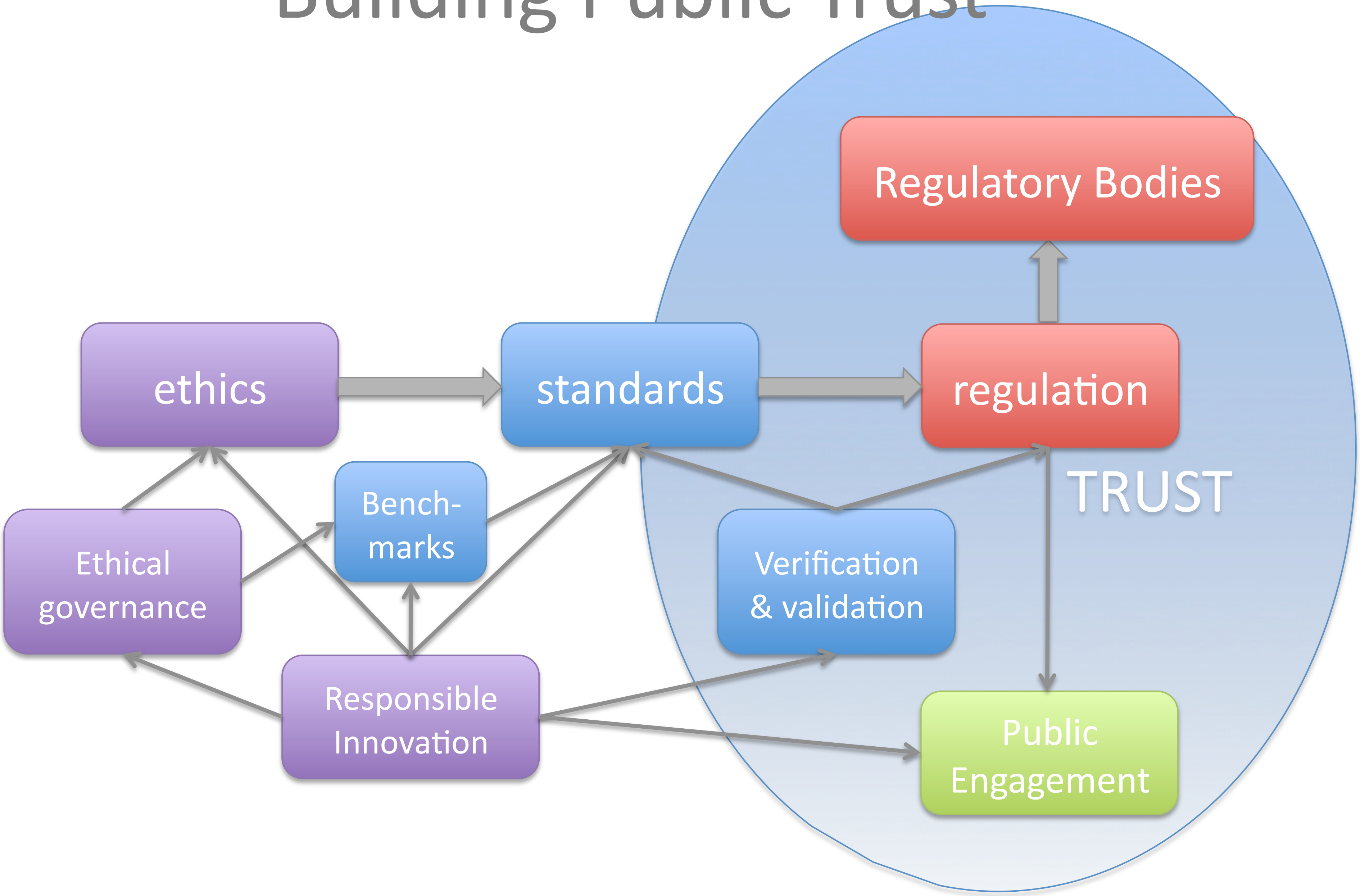
Responsible Innovation

Open science
reproducibility

Design
for Safety

Rome declaration on RRI
Foundation for Responsible Robotics
EPSRC ORBIT

Building Public Trust



BS 8611:2016

BSI Standards Publication

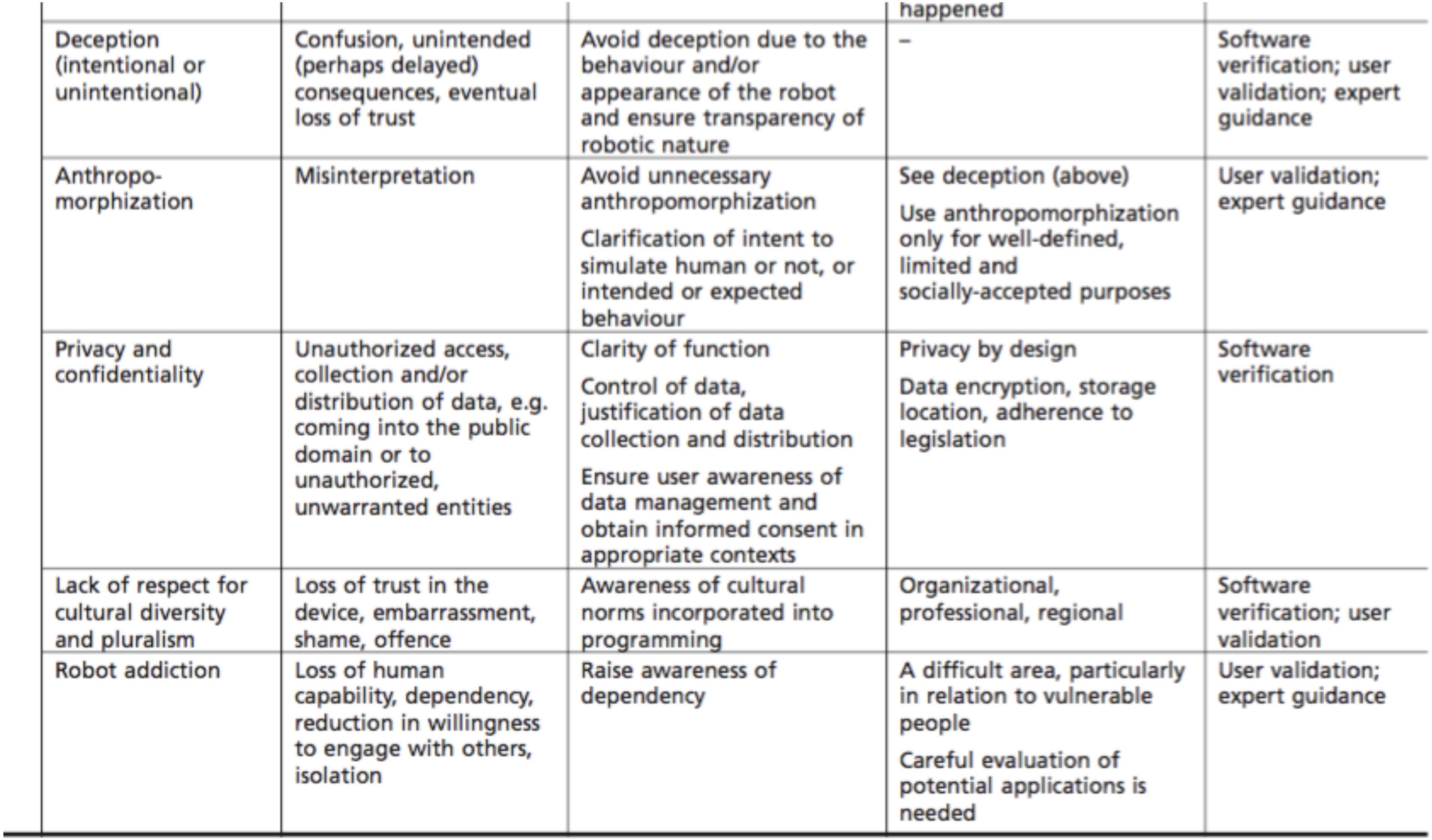
Robots and robotic devices

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

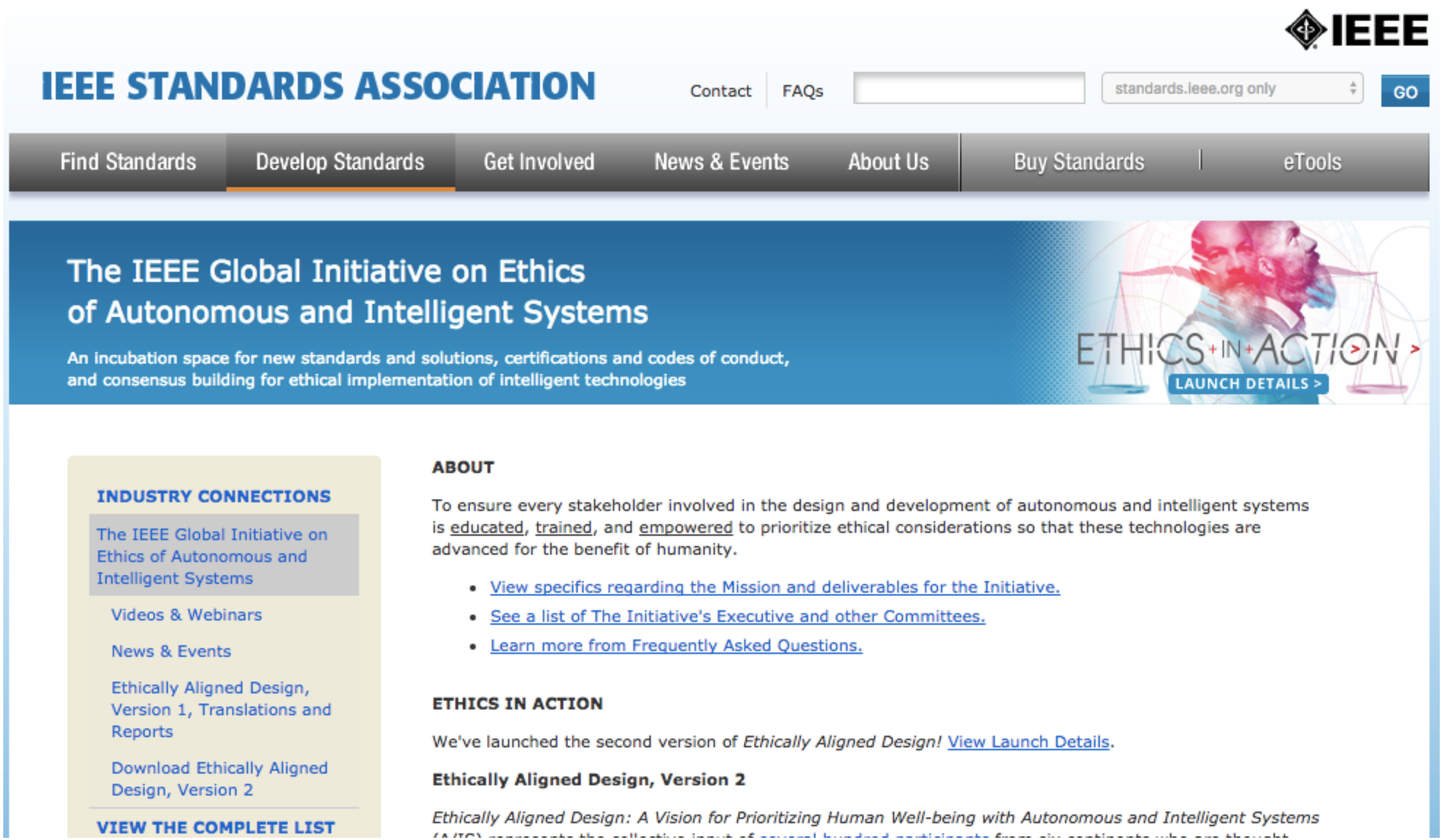
Ethical Risk Assessment

- BS8611 is a set of 20 distinct *ethical hazards and risks*, grouped under four categories:
  - societal,
  - application,
  - commercial/financial, and
  - 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
Anthropomorphization	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



The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems

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

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

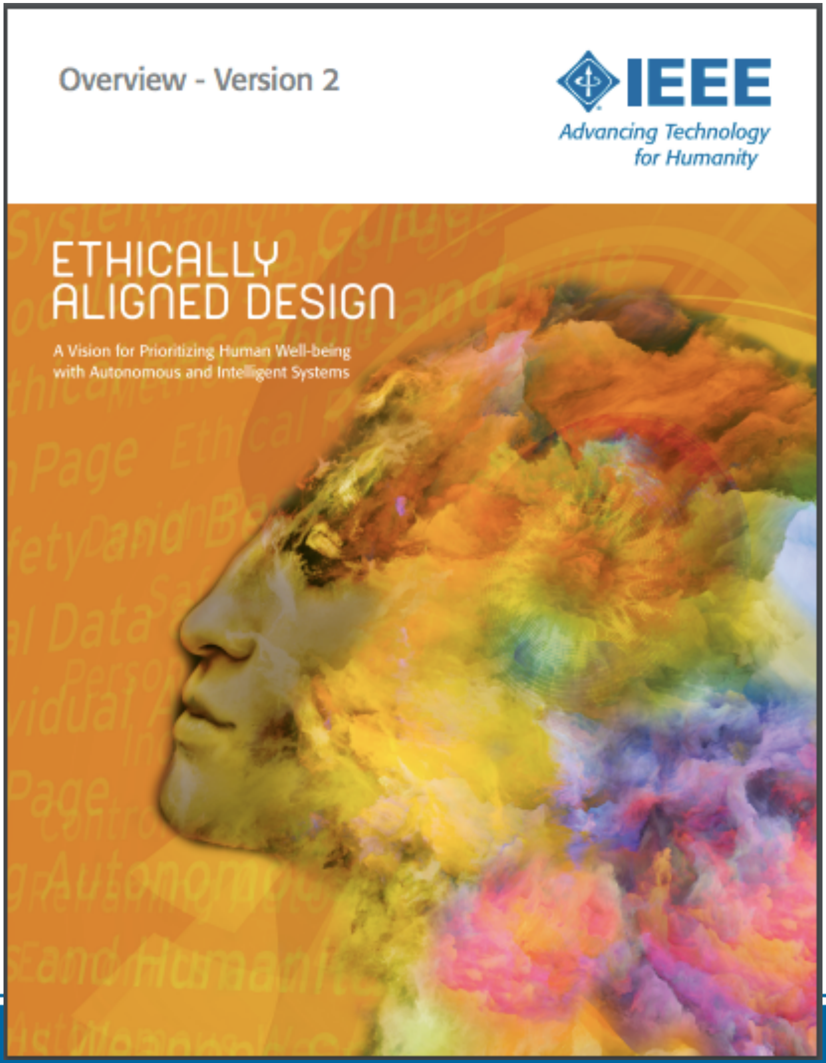
We've launched the second version of *Ethically Aligned Design!* View Launch Details.

Ethically Aligned Design, Version 2

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

http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html

Deliverables

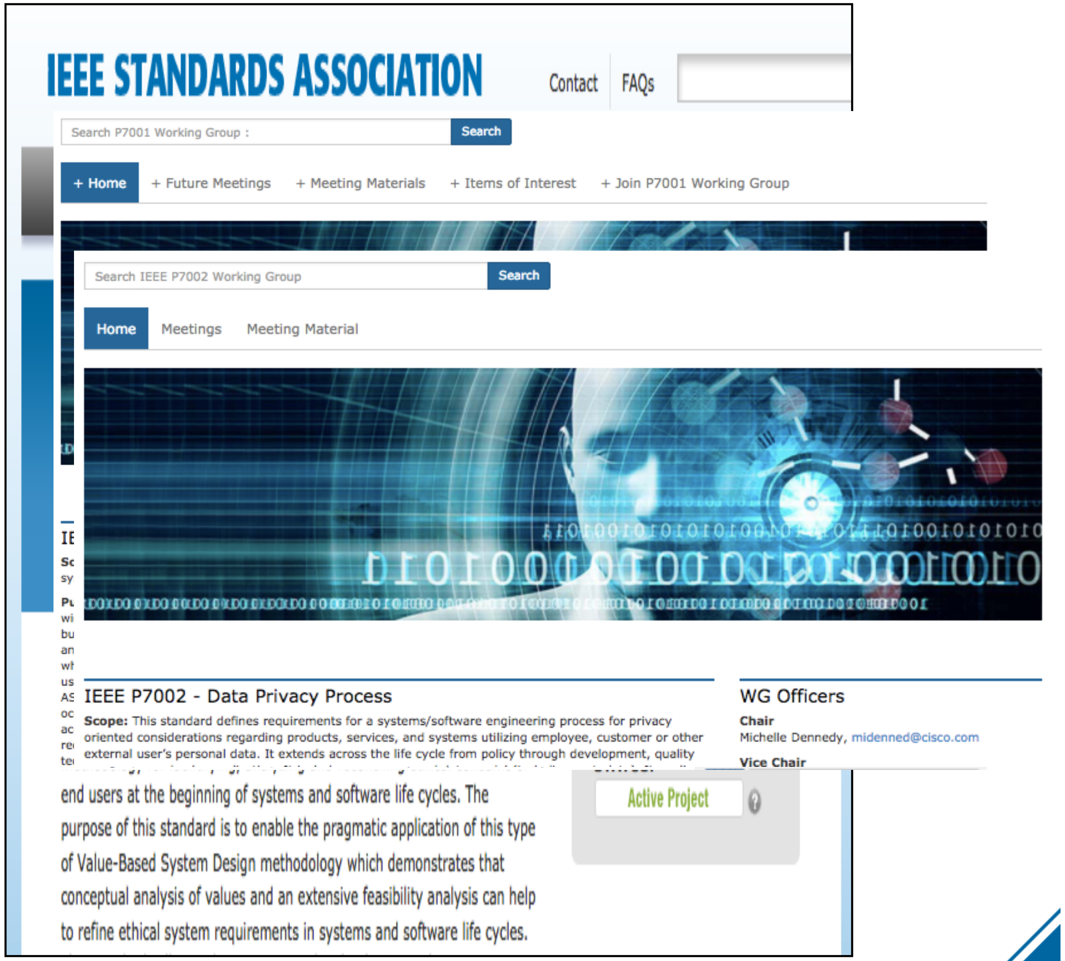


Overview - Version 2

IEEE Advancing Technology for Humanity

ETHICALLY ALIGNED DESIGN

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



IEEE P7002 - Data Privacy Process

Scope: This standard defines requirements for a systems/software engineering process for privacy oriented considerations regarding products, services, and systems utilizing employee, customer or other external user's personal data. It extends across the life cycle from policy through development, quality

WG Officers

Chair

Michelle Dennedy, midenned@cisco.com

Vice Chair

end users at the beginning of systems and software life cycles. The purpose of this standard is to enable the pragmatic application of this type of Value-Based System Design methodology which demonstrates that conceptual analysis of values and an extensive feasibility analysis can help to refine ethical system requirements in systems and software life cycles.

Human standards in draft 1

- **P7000** — **Model Process** for Addressing Ethical Concerns during System Design
  - http://standards.ieee.org/develop/project/7000.html
  - Aims to establish a value-based system design methodology
- **P7001** — **Transparency** of Autonomous Systems
  - http://standards.ieee.org/develop/project/7001.html
  - Aims to set out measurable, testable levels of transparency for a range of different stakeholders
- **P7002** — **Data Privacy** Process
  - http://standards.ieee.org/develop/project/7002.html
  - Aims to create one overall methodological approach that specifies practices to manage privacy issues
- **P7003** — **Algorithmic Bias** Considerations
  - http://standards.ieee.org/develop/project/7003.html
  - Aims to specify methodologies to ensure that negative bias in algorithms has been addressed and eliminated

Human standards in draft 2

- **IEEE P7004** Standard on **Child and Student Data Governance**
- **IEEE P7005** Standard on E**mployer Data Governance**
- **IEEE P7006** Standard on **Personal Data** AI Agent Working Group
- **IEEE P7007** **Ontological** Standard for Ethically driven Robotics and Automation Systems
- **IEEE P7008** Standard for Ethically Driven **Nudging** for Robotic, Intelligent and Autonomous Systems .
- **IEEE P7009** Standard for **Fail-Safe Design** of Autonomous and Semi-Autonomous Systems
- **IEEE P7010** **Wellbeing Metrics** Standard for Ethical Artificial Intelligence and Autonomous Systems

P7001 - Transparency

- Based on the principle that it should always be possible to discover why an autonomous system made a particular decision
- Transparency is not one thing
- Stakeholders:
  - Users
  - Safety testers/certifiers
  - Accident investigators
  - Lawyers/expert witnesses
  - The public at large

Transparency

- 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 social-technical systems: they are designed to work 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 isn’t 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.
- Who are the stakeholders?
  - Users
  - Safety certification engineers or agencies
  - Accident investigators
  - Lawyers or expert witnesses
  - Wider society

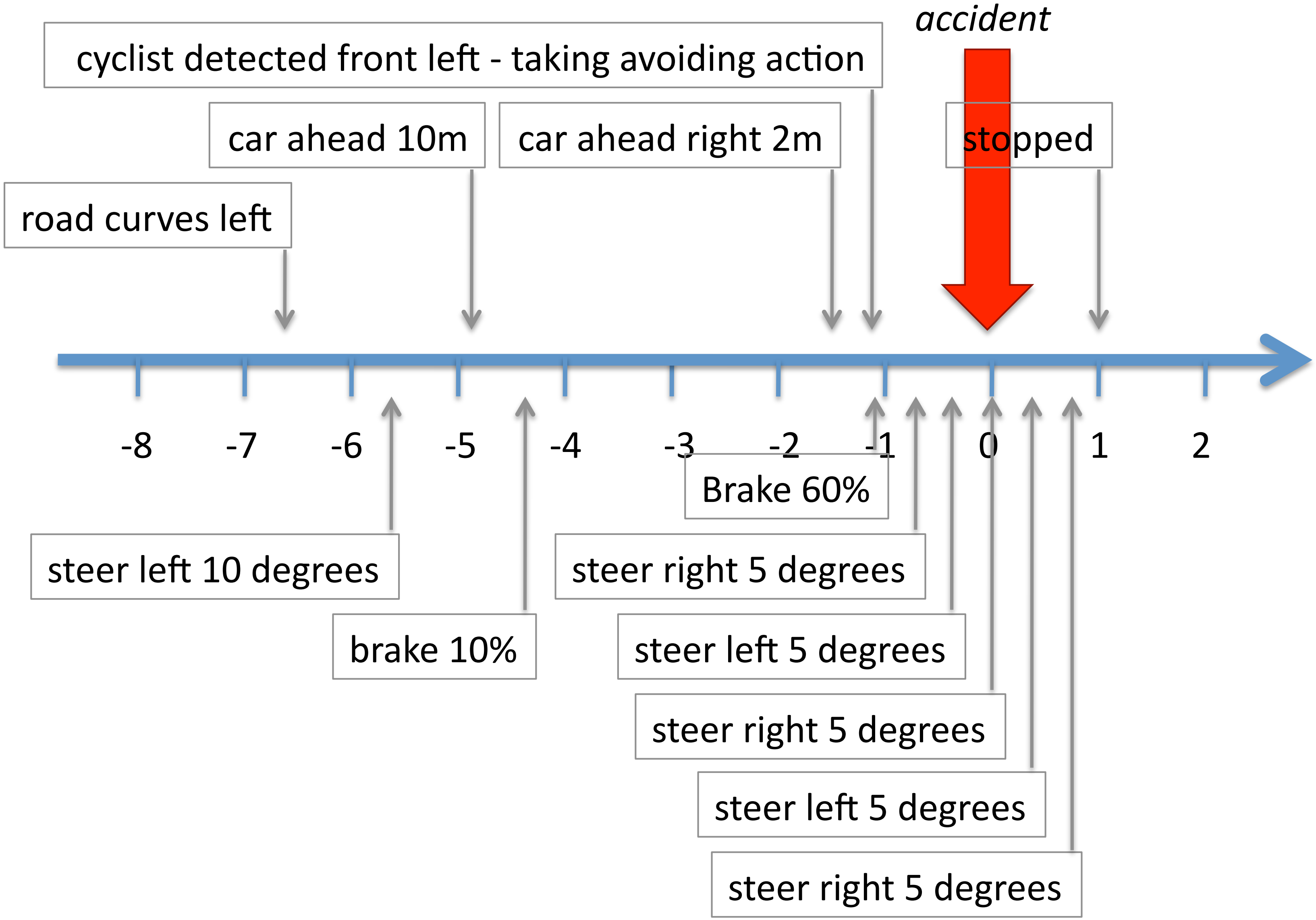
Transparency for Accident Investigators

- 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).

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

An annotated timeline



A human process

“Investigation begins on robot security after child is hurt”
CNCB News, July 2016

A proliferation of principles

- Asimov's three laws of Robotics (1950)
- EPSRC/AHRC Principles of Robotics (2010)
- Future of Life Institute Asilomar principles for beneficial AI (Jan 2017)
- The ACM US Public Policy Council Principles for Algorithmic Transparency and Accountability (Jan 2017)
- Japanese Society for Artificial Intelligence (JSAI) Ethical Guidelines (Feb 2017)
- Draft principles of The Future Society's Science, Law and Society Initiative (Oct 2017)
- Montréal Declaration for Responsible AI draft principles (Nov 2017)
- IEEE General Principles of Ethical Autonomous and Intelligent Systems (Dec 2017)
- UNI Global Union Top 10 Principles for Ethical AI (Dec 2017)

What is ethical governance (and who's doing it?*)

- Have an ethical code of conduct.
  - so that everyone in the organisation understands what is expected of them. And provide a mechanism for whistleblowers.
- Provide ethics training for everyone, without exception
  - Ethics, like quality, is not something you can do as as add-on; simply appointing an ethics manager, while not a bad idea, is not enough.
- Undertake ethical risk assessments of all new products, and act upon the findings of those assessments.
- Be transparent about your ethical governance.
  - Of course your robots and AIs must be transparent too, but here I mean *transparency of process*, not product.
- Really *value* ethical governance.

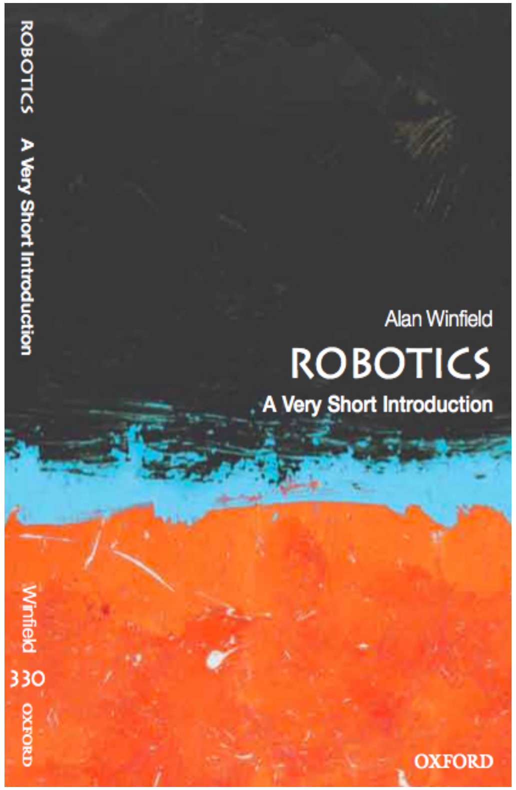
brl
Bristol Robotics Laboratory

https://alanwinfield.blogspot.com/2018/02/ethical-governance-what-is-it-and-whos.html

Thoughts and Questions

- What kind of governance do we want/need?
- What kind of regulatory bodies?
  - An equivalent of the European Aviation Safety Authority (EASA) for driverless cars..?
- What standards are missing?
- Who's working on these questions…?

Winfield AFT, Jirotka M. 2018 Ethical governance is essential to building trust in robotics and artificial intelligence systems. Phil. Trans. R. Soc. A 376: 20180085.



Thank you!