brl
Bristol Robotics Laboratory

UWE BRISTOL
University of the West of England

University of BRISTOL

Responsible AI

From principles to practice: a Toolkit for AI ethicists

Alan FT Winfield
Bristol Robotics Laboratory
Science Communication Unit
Twitter: @alan_winfield

“Exploring our AI Potential”
UK – Romania Conference
Bucharest
28 February 2024

Outline

I. Introducing AI Ethics

II. Why AI ethics are so important

III. Linking ethics to standards and regulation

IV. Robot/AI accident investigation

V. Tools and Methods for Responsible AI

brl
Bristol Robotics Laboratory

Part I Introducing robot/AI ethics

brl
Bristol Robotics Laboratory

Ethics

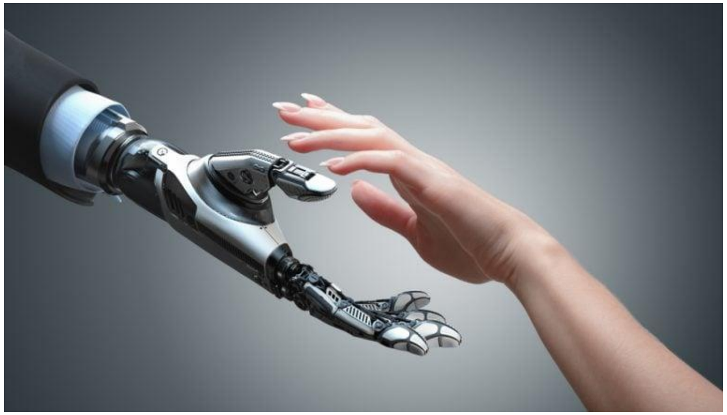
- Ethics can be viewed as moral principles which govern a person's behaviour or the conduct of a group

shutterstock.com • 100605031

"Treat others how you wish to be treated"

Robot/AI ethics

- Robot/AI ethics is concerned with *the ethical impact of robots/AIs, on individuals, society and the environment*, and how any negative impacts can be mitigated



SociallyResponsible
ResponsibleInnovation
Explainable
RespectingDiversity
HumanRights
Transparent
AIEthics
Sustainable
Dignity
Privacy
Moral
Unbiased
NonGendered
Accountable

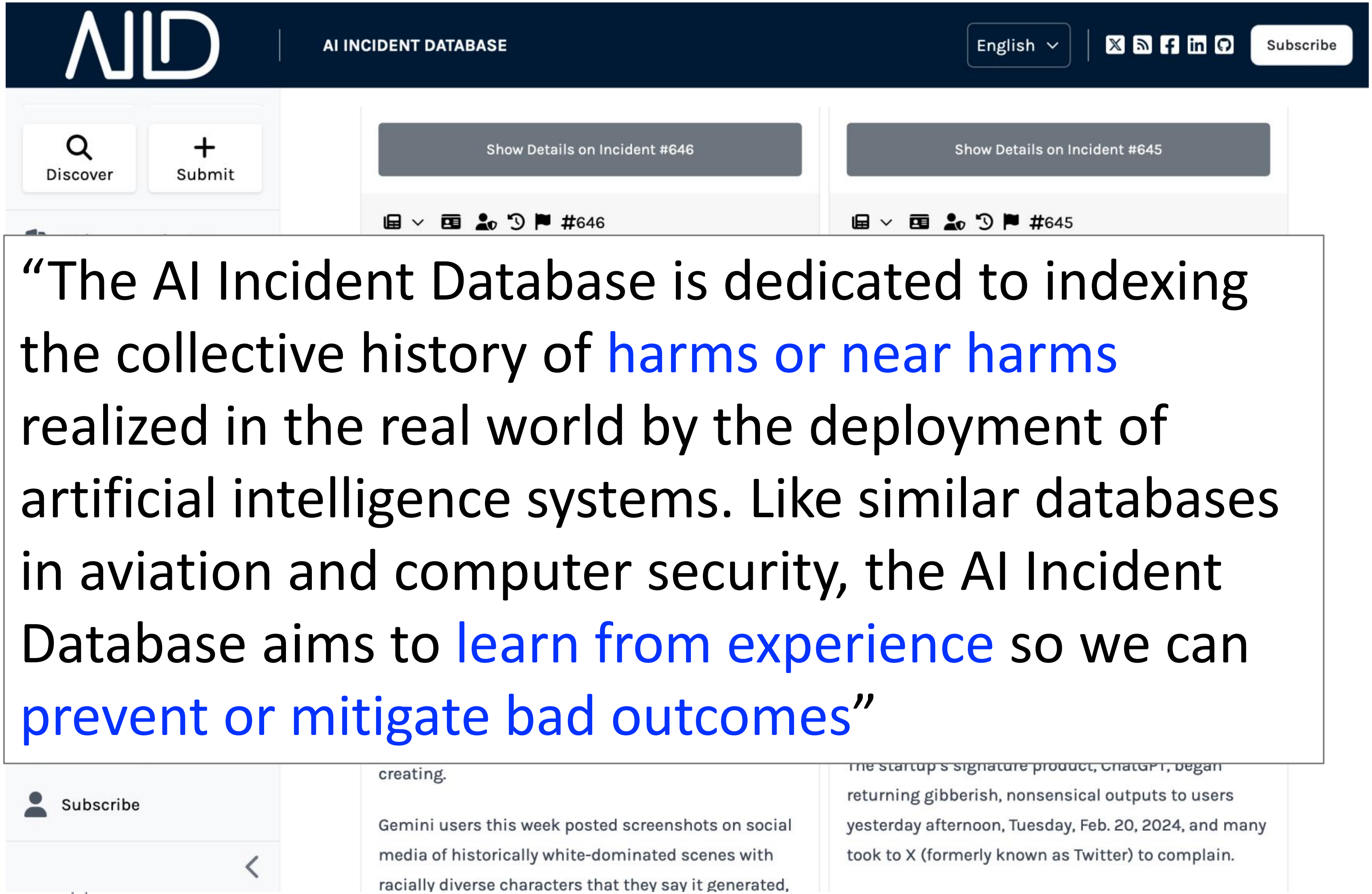
So, what do roboethicists do?

- In practice robot/AI ethicists:
  - Develop *ethical principles* in order to support and guide the way in which robots are designed and deployed in society
  - Develop methods for *Values Driven Design*
  - Draft new *Ethical Standards* in robotics and AI
  - *Advocate* for the adoption of these principles, methods and standards and, if necessary, new regulation

Part II Why AI Ethics are so important

brl
Bristol Robotics Laboratory

The AI Incidents Database



creating.

Gemini users this week posted screenshots on social media of historically white-dominated scenes with racially diverse characters that they say it generated,

The startup's signature product, ChatGPT, began returning gibberish, nonsensical outputs to users yesterday afternoon, Tuesday, Feb. 20, 2024, and many took to X (formerly known as Twitter) to complain.

"The AI Incident Database is dedicated to indexing the collective history of harms or near harms realized in the real world by the deployment of artificial intelligence systems. Like similar databases in aviation and computer security, the AI Incident Database aims to learn from experience so we can prevent or mitigate bad outcomes"

https://incidentdatabase.ai/

How can robots and AIs cause harm?

- Physical harm: Robot/AI accidents
- Psychological harms: deception, over dependancy, over trusting
- Societal harms: loss of privacy, loss of jobs
- Environmental: high energy costs, unrepairable or unrecyclable tech.

AI RISKS	HARMS
Discrimination, Exclusion and Toxicity	justified offense, material (allocational) harm, and the unjust representation or treatment of marginalized groups
Information Hazards	privacy violations and safety risks
Misinformation Harms	deception, material harm, or unethical actions by humans who take the AI prediction to be factually correct, as well as wider societal distrust in shared information
Malicious Uses	undermining public discourse, crimes such as fraud, personalized disinformation campaigns, and the weaponization or production of malicious code
Human-Computer Interaction Harms	unsafe use due to users misjudging or mistakenly trusting the AI, psychological vulnerabilities and privacy violations of the user, and social harm from perpetuating discriminatory associations via product design (e.g. making “assistant” tools by default “female.”)
Automation, access, and environmental harms	increasing social inequalities from uneven distribution of risk and benefits, loss of high-quality and safe employment, and environmental harm

brl
Bristol Robotics Laboratory

Extract from Table 1 in: Weidinger et al (2021) Ethical and social risks of harm from Language Models, https://arxiv.org/abs/2112.04359

Part III Linking ethics to standards and regulation

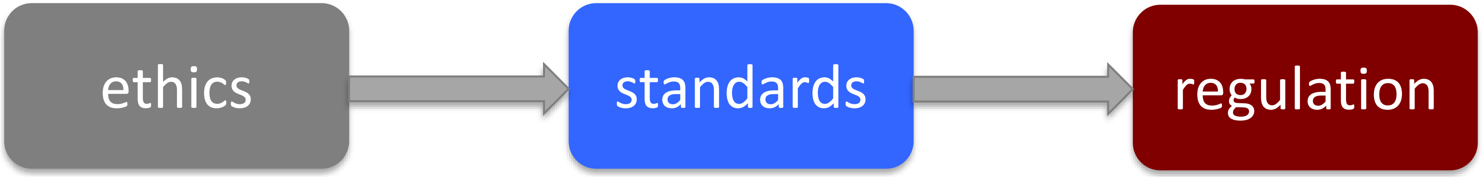
brl
Bristol Robotics Laboratory

From ethical principles to ethical standards and regulation*

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

Emerging ethical standards:
BS 8611:2016
IEEE P700X

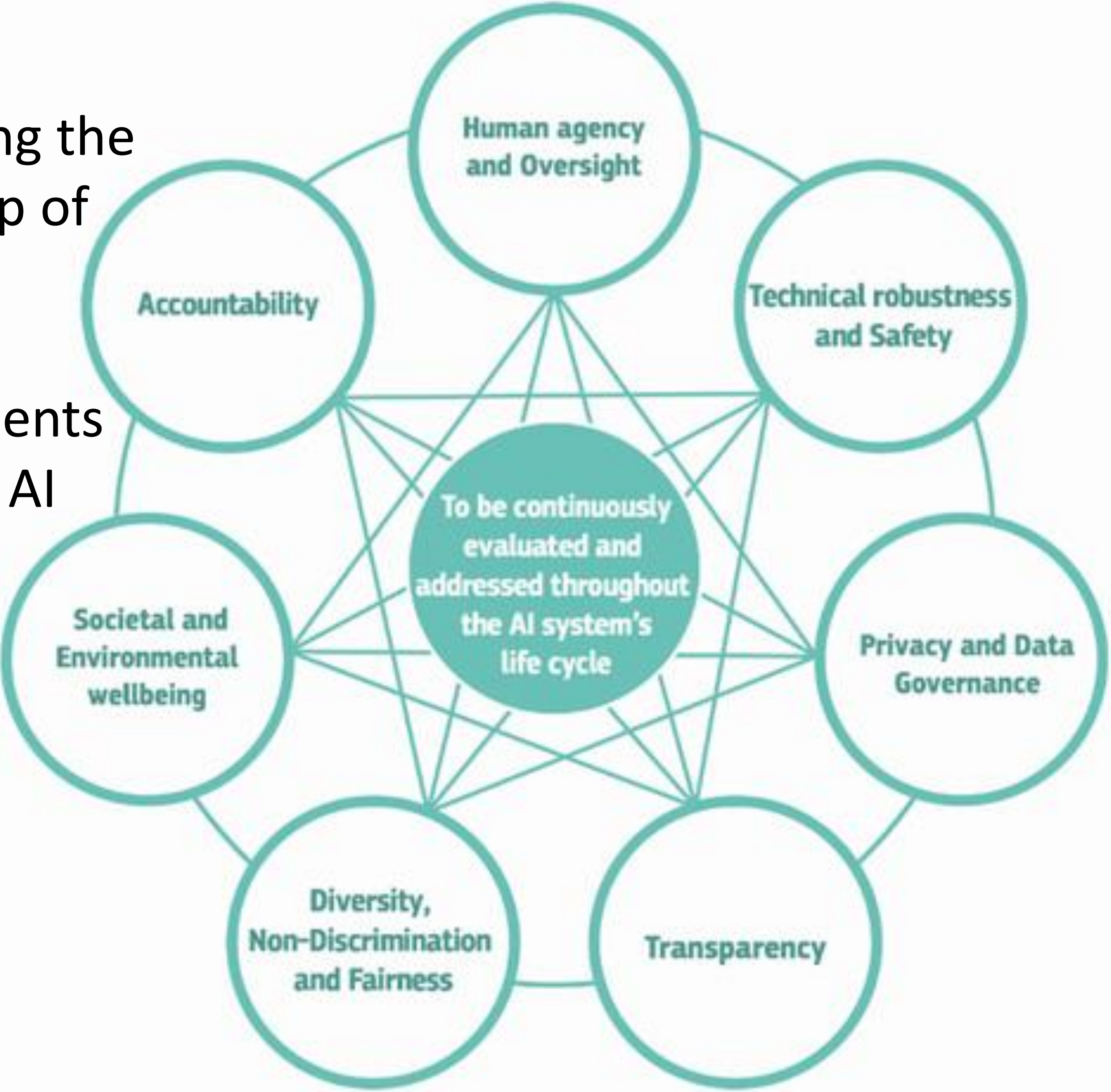
Emerging regulation:
Driverless cars?
Assistive robotics?
AIs?



*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

Diagram showing the interrelationship of the European Commission’s seven requirements for trustworthy AI (2019)

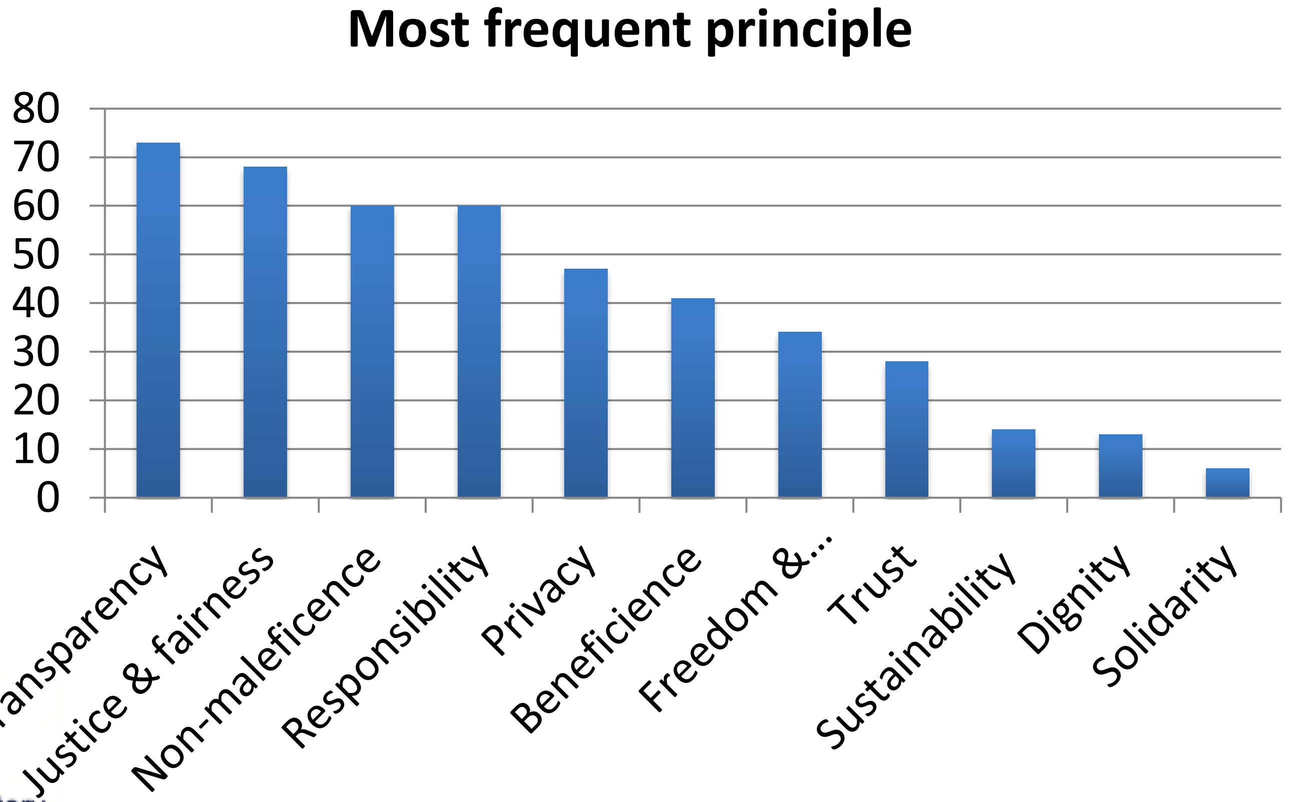
European Commission


Human agency and Oversight
Technical robustness and Safety
Privacy and Data Governance
Transparency
Diversity, Non-Discrimination and Fairness
Societal and Environmental wellbeing
Accountability
To be continuously evaluated and addressed throughout the AI system’s life cycle

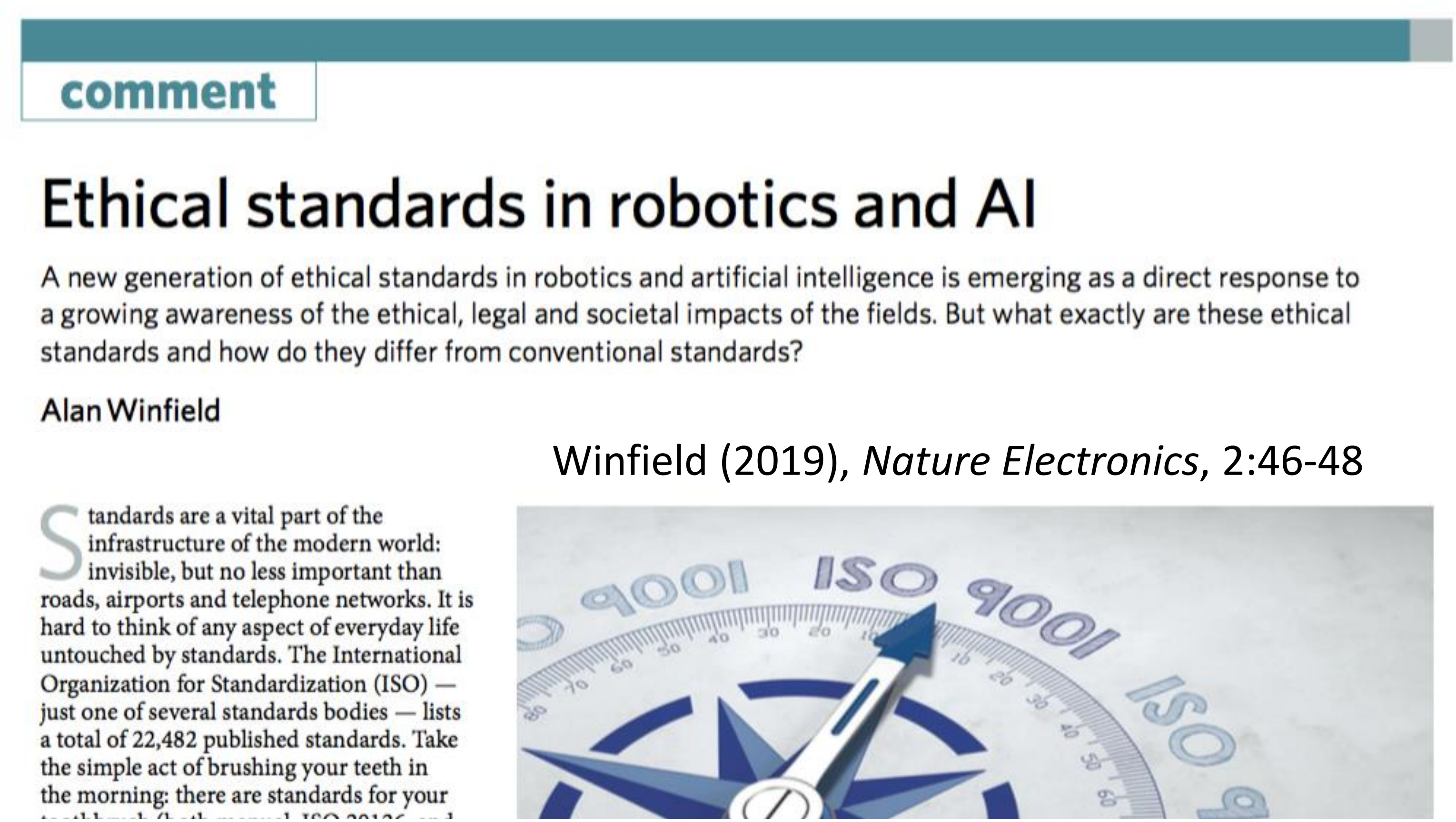
brl
Bristol Robotics Laborat

Comparing ethical frameworks

- A very comprehensive survey and analysis of 84 ethical frameworks in AI (Jobin *et al* 2019, Nature Machine Intellligence) lists the principles which appear most often in ethical frameworks for AI.



Standards are infrastructure



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

Winfield (2019), *Nature Electronics*, 2:46-48

brl
Bristol Robot

Ethical Risk Assessment

BS 8611:2016

BSI Standards Publication

Robots and robotic devices

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

RoboTed: a worked example

- Consider a fictional robot toy called RoboTed.
  - Physical risks
  - Psychological risks
  - Privacy & security risks
  - Environmental risks

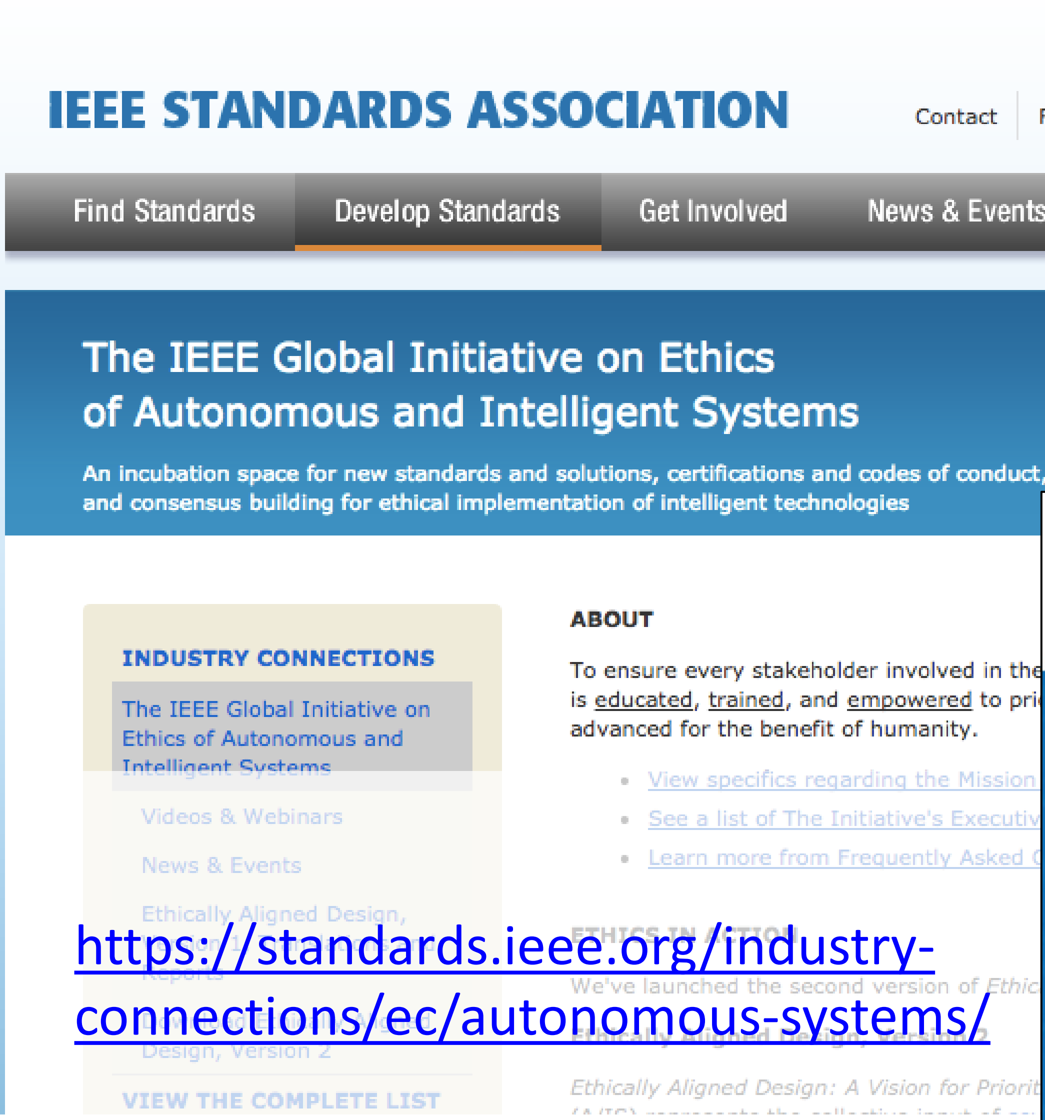
Winfield AFT and Winkle K (2020) RoboTed: a case study in Ethical Risk Assessment, 5th Int. Conf. on Robot Ethics and Standards (ICRES 2020), 28-29 September 2020. [arXiv preprint: 2007.15864]

brl

Bristol Robotics Laboratory

RoboTed 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



IEEE

IEEE STANDARDS ASSOCIATION

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

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

Download Ethically Aligned Design, Version 2

VIEW THE COMPLETE LIST

ABOUT

To ensure every stakeholder involved in the
is educated, trained, and empowered to pri
advanced for the benefit of humanity.

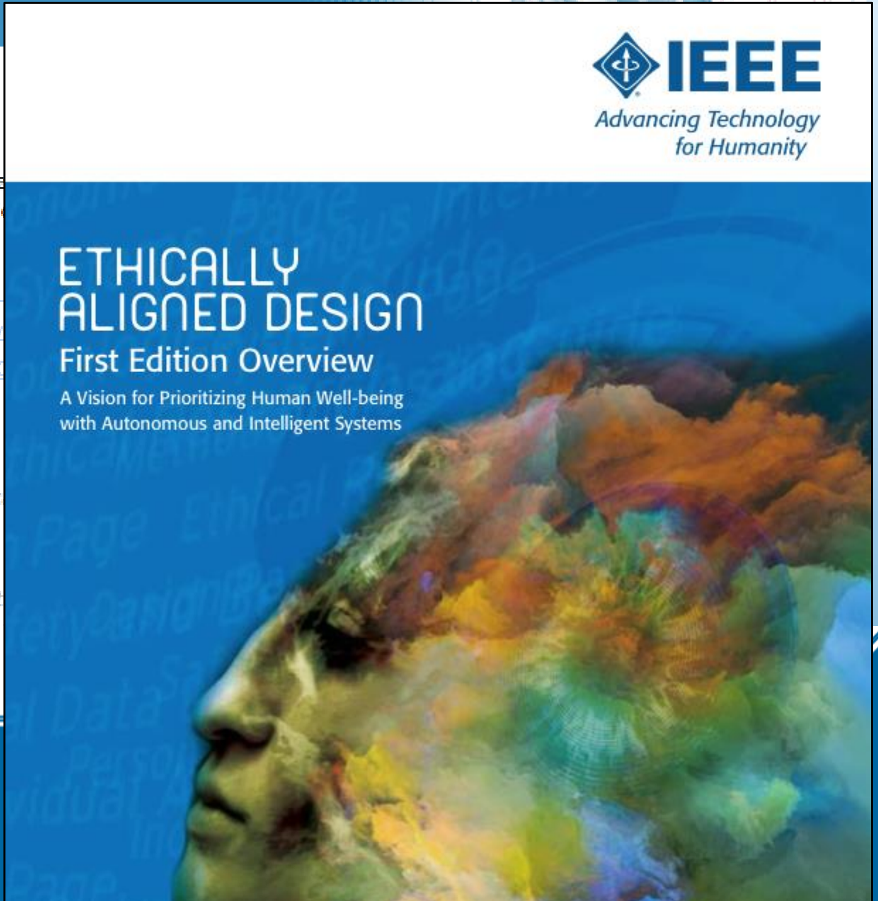
- View specifics regarding the Mission
- See a list of The Initiative's Executiv
- Learn more from Frequently Asked C

ETHICS IN ACTION

We've launched the second version of *Ethic*

Ethically Aligned Design, Version 2

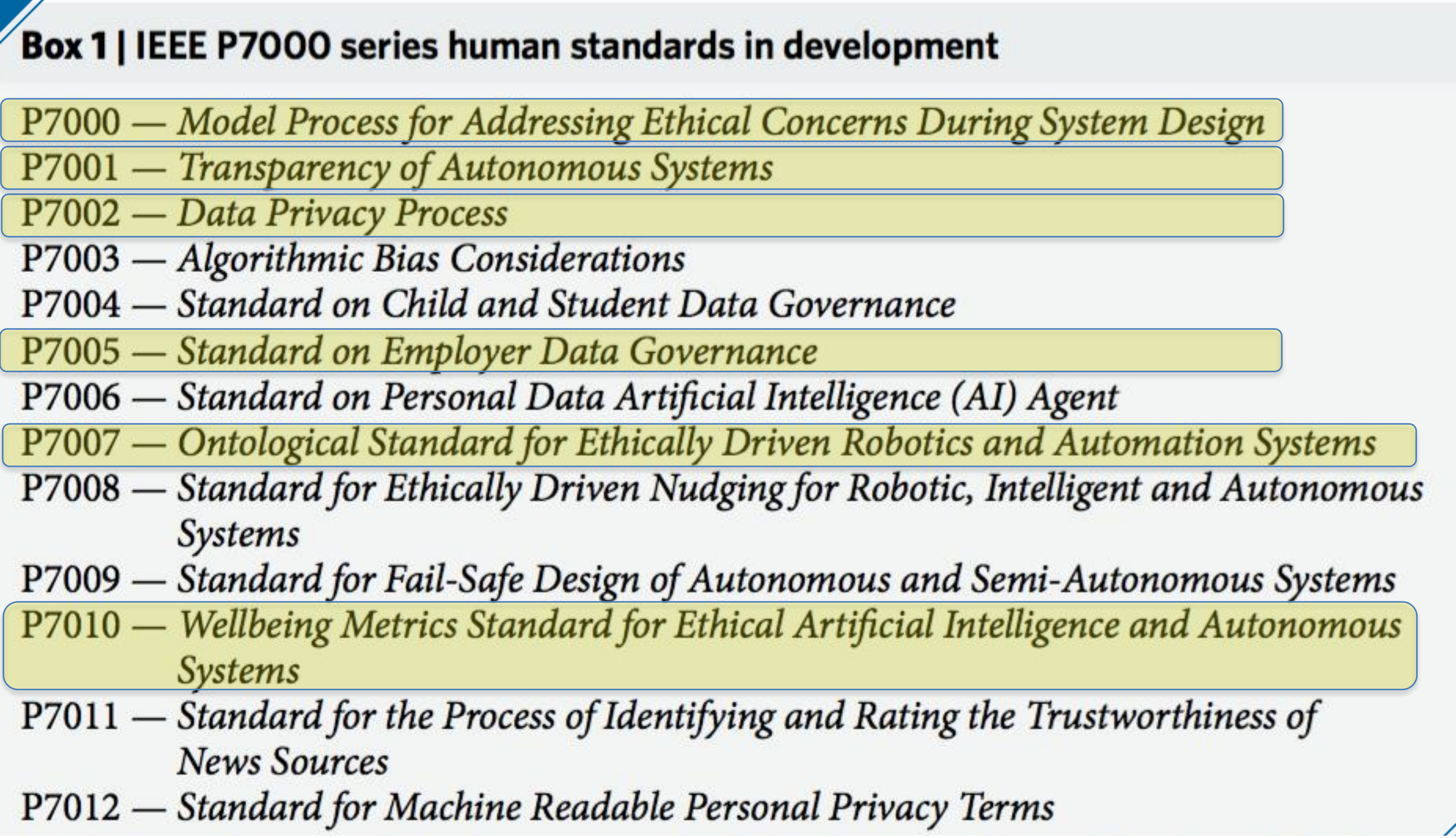
Ethically Aligned Design: A Vision for Priorit



IEEE
Advancing Technology for Humanity

ETHICALLY ALIGNED DESIGN
First Edition Overview
A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems

https://standards.ieee.org/industry-connections/ec/autonomous-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 - *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 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

https://ieeexplore.ieee.org/document/9726144

brl
Bristol Robotics Laboratory

Why is transparency so important?

- Autonomous Systems 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
- *All* 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 relied upon.

Bristol Robotics Laboratory

7001 Processes

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

Winfield *et al* (2021) IEEE P7001: A Proposed Standard on Transparency, Frontiers in Robotics and AI, 8
https://www.frontiersin.org/articles/10.3389/frobt.2021.665729/full

Part IV Robot/AI accident investigation

UNIVERSITY OF OXFORD

ROBOTIPS

RESPONSIBLE ROBOTS FOR THE DIGITAL ECONOMY

brl

Bristol Robotics Laboratory

www.robotips.co.uk

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

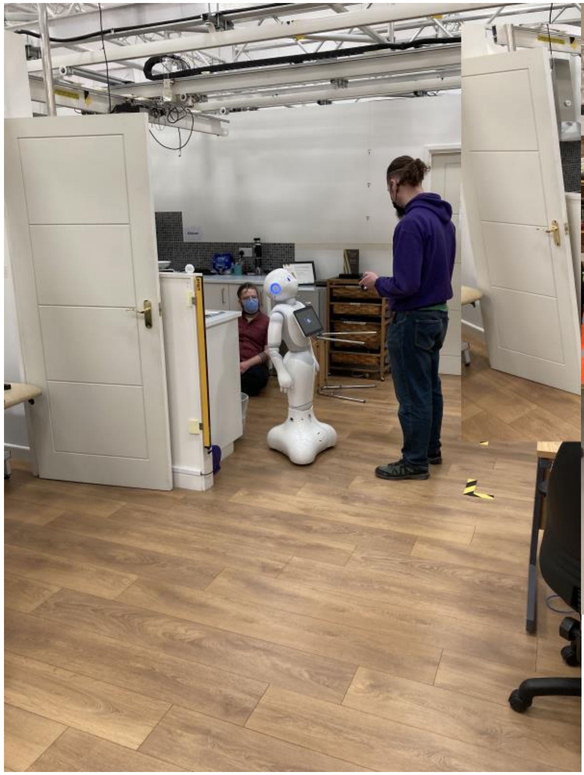
brl
Bristol Robotics Laboratory

A human process

In RoboTIPS we are running a series of simulated accident scenarios, with human volunteers as:

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

Physical accident involving an Assisted Living Robot, Bristol



Bristol Robotics Laboratory
Personalised Assisted Living Studio

Smart toy robot that caused psychological harm, Oxford

brl
Bristol Robotics Laboratory

Part V Practice: Tools and Methods for Responsible AI

Bristol Robotics Laboratory

Tools and methods for Responsible AI

- Ethical Risk Assessment
  - An extremely powerful method which aims to anticipate and mitigate risks, and hence head off problems before they occur
- Design for Transparency
  - IEEE Std 7001-2021 provides tools for measuring and testing transparency and – if necessary – raising the levels of transparency in a system
- Accident/Incident investigation
  - All accident investigation aims to answer three questions: *what* happened, *why,* and *how* the system can be improved to ensure it cannot happen again

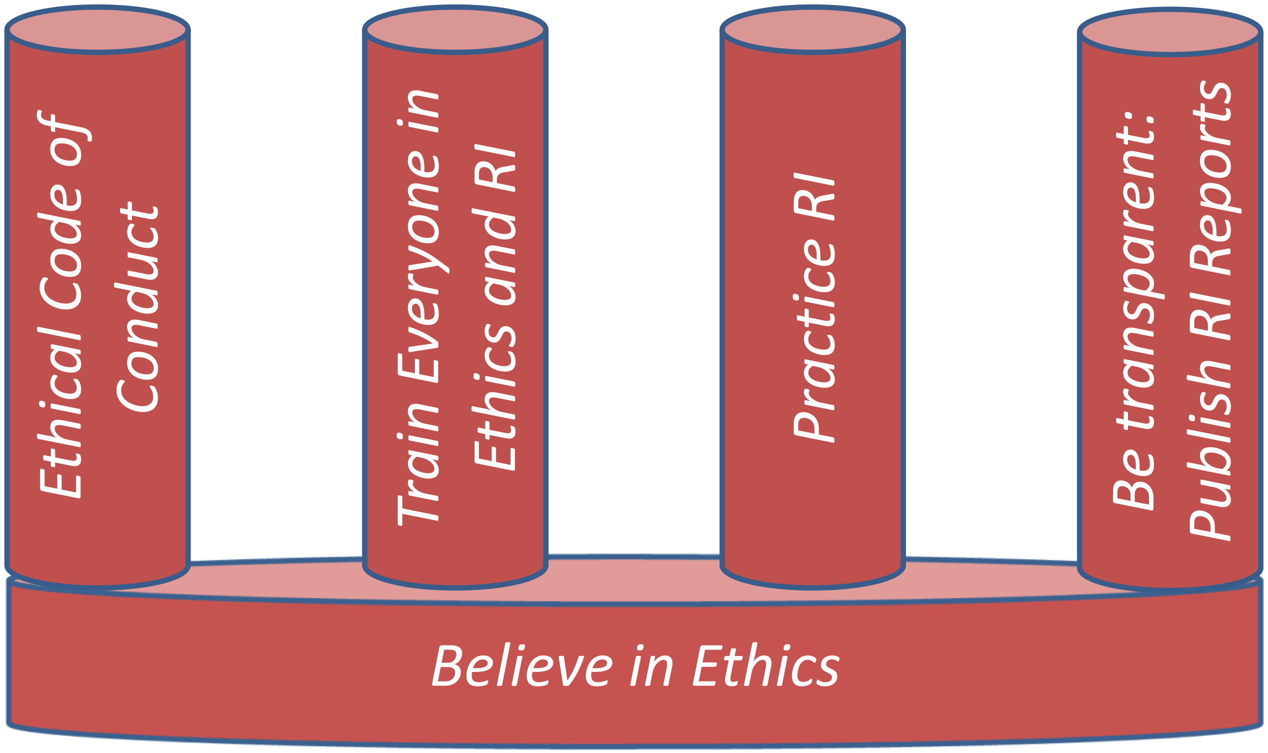
A framework for Independent Audit

- In a paper published in 2021 we set out a framework called the Independent Audit of AI Systems (IAAIS – or 'eyes'), with three components:
    - Prospective Assessments before AI systems are implemented (i.e. ethical risk assessment)
    - Maintenance of an Audit trail to analyze failures and help assess accountability
    - Ensure system Adherence to regulatory requirements, through independent oversight boards

Bristol Robotics Laboratory

Falco et al (2021) Governing AI Safety through Independent Audits, Nature Machine Intelligence, https://www.nature.com/articles/s42256-021-00370-7

The Five pillars of Ethical Governance



Ethical Code of Conduct

Train Everyone in Ethics and RI

Practice RI

Be transparent: Publish RI Reports

Believe in Ethics

Winfield and Jirotka (2018), Ethical Governance in Robotics and AI, Phil. Trans. R. Soc. A 376: 20180085

Regulation and related standards in AI

- AI Regulation
  - The EU AI Act aims to ensure “AI systems used in the EU are safe, transparent, traceable, non-discriminatory and environmentally friendly”
- AI Standards drafting is underway *at pace*
  - IEEE SA published six 700X AI standards, more in draft
  - ISO/IEC joint technical committee 1/SC 42 on AI has published 20 AI standards since 2018, and a further 30 are in development.

Bristol Robotics Laboratory

Winfield and Studley (2024) On the relationship between Benchmarking, Standards and Certification in Robotics and AI, in press. Preprint https://arxiv.org/abs/2309.12139

Thank you!

Thanks to the RoboTIPS team led by Prof Marina Jirotka

Take home messages:

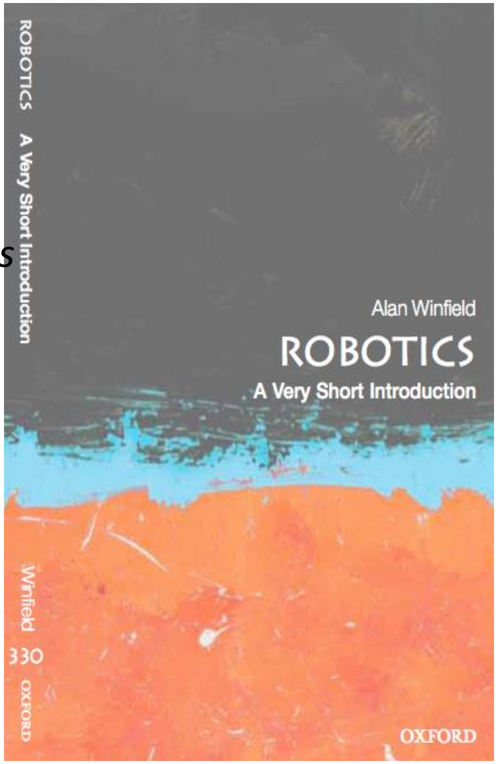
- We should all be *very* worried by the actual (and potential) harms being caused by AI
- Powerful Tools and Methods *already exist* for Responsible AI.

EPSRC

Engineering and Physical Sciences Research Council

Key references:

- Winfield and Jirotka (2018), Ethical Governance in Robotics and AI, *Phil. Trans. R. Soc.* A 376: 20180085
- Winfield (2019) Ethical standards in Robotics and AI. *Nature Electronics* 2(2) 46-48.
- Winfield and Jirotka (2017) The case for an ethical black box, Towards Autonomous Robotic Systems (TAROS), LNCS 10454, 262-273
- Winfield *et al* (2021), Robot Accident Investigation: a case study in responsible robotics, Software engineering for robotics, 165-187.
- Winfield et al (2021) IEEE P7001: A Proposed Standard on Transparency. *Front. Robot. AI* 8:665729. doi: 10.3389/frobt.2021.665729



Bristol Robotics Laboratory