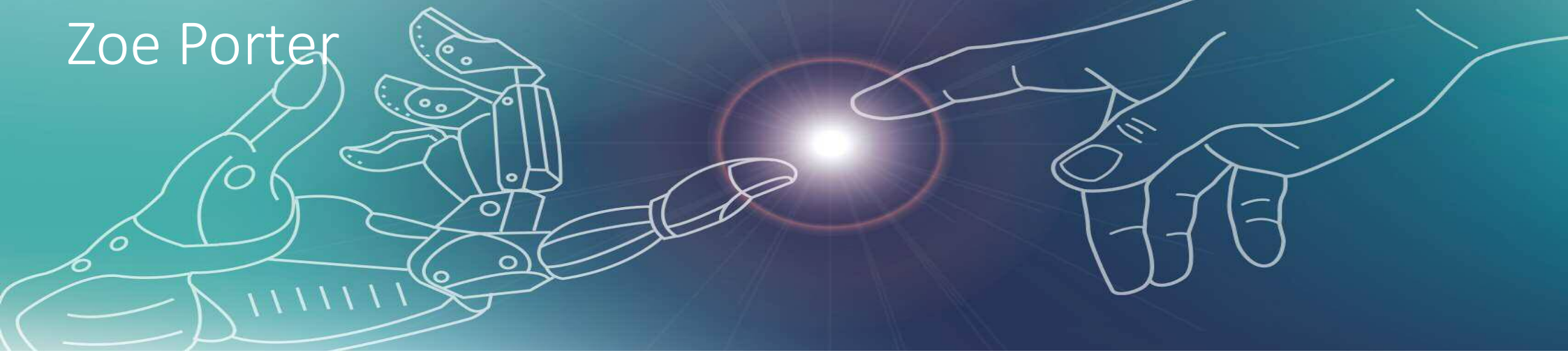
UNIVERSITY of York

ASSURING AUTONOMY
INTERNATIONAL PROGRAMME

Beyond safety assurance: assuring the ethics of AI and autonomous systems

Zoe Porter

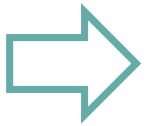


Safety Assurance

Assuring Autonomy International Programme (AAIP)

£12m partnership between Lloyd's Register Foundation and the University of York.

- Academic research
- Demonstrator projects
- Empirical evaluation

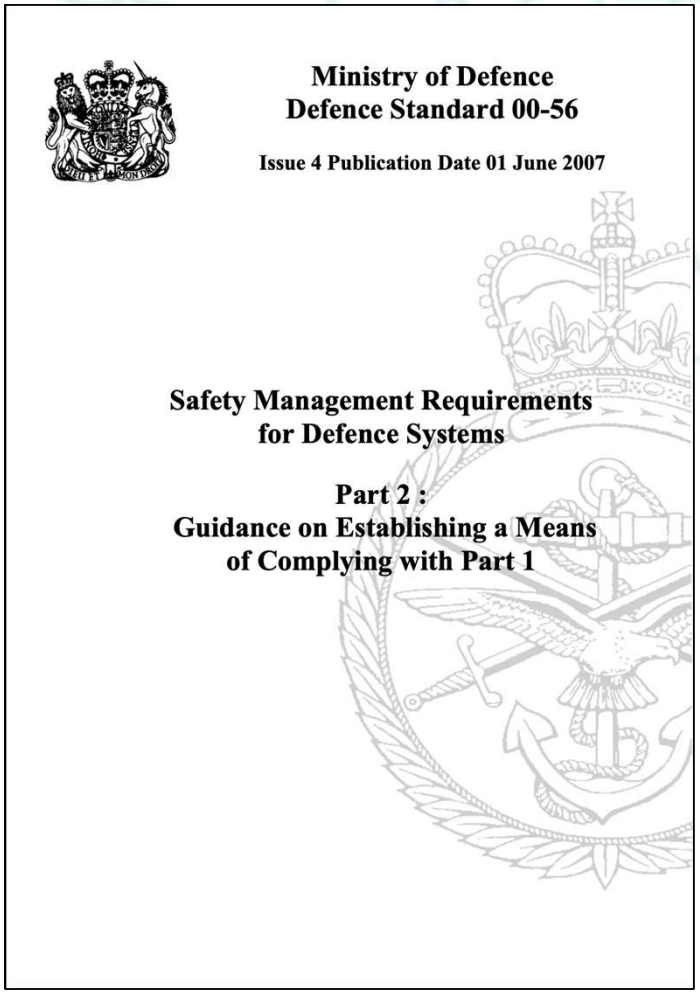


- Free, expert guidance on safety assurance
- Bespoke CPD and online training opportunities

Safety Cases

UK Defence Standard 00:56

*A **structured argument**, supported by a **body of evidence** that provides a compelling, comprehensible and valid case that a system is safe for a given application in a given operating environment.*



Ministry of Defence
Defence Standard 00-56

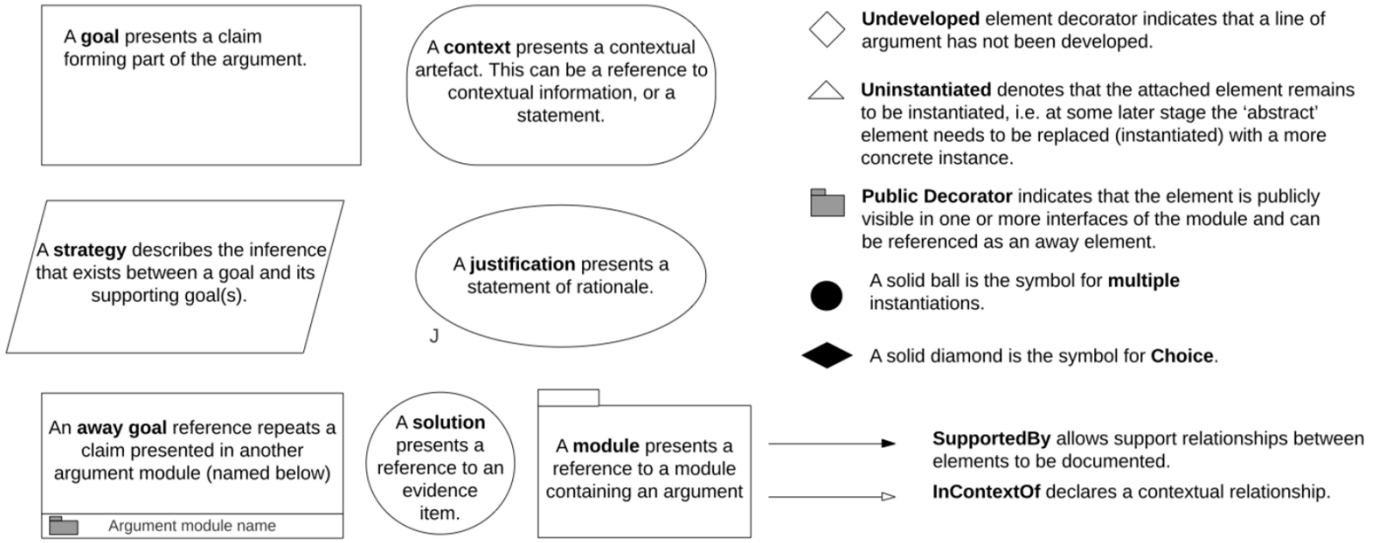
Issue 4 Publication Date 01 June 2007

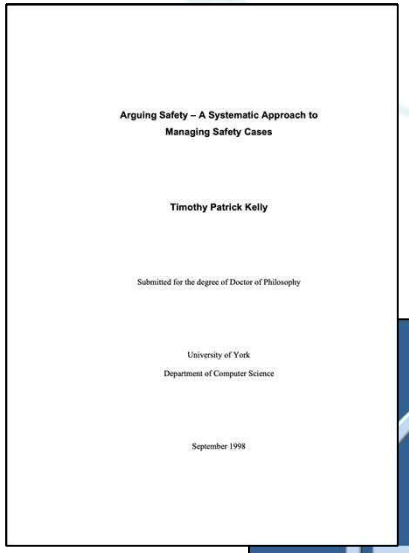
Safety Management Requirements
for Defence Systems

Part 2 :
Guidance on Establishing a Means
of Complying with Part 1

Goal-structuring notation (GSN)

A graphical approach to presenting the structure of a safety argument



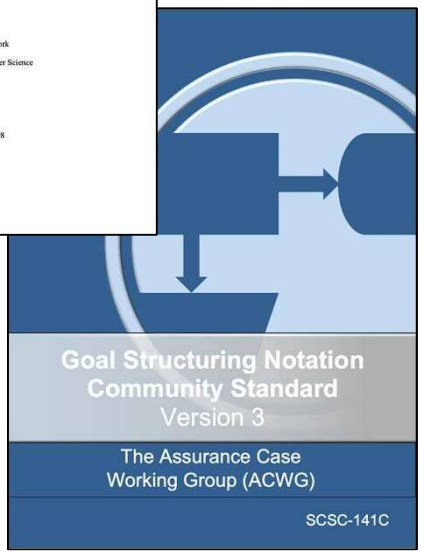

Arguing Safety – A Systematic Approach to Managing Safety Cases

Timothy Patrick Kelly

Submitted for the degree of Doctor of Philosophy

University of York
Department of Computer Science

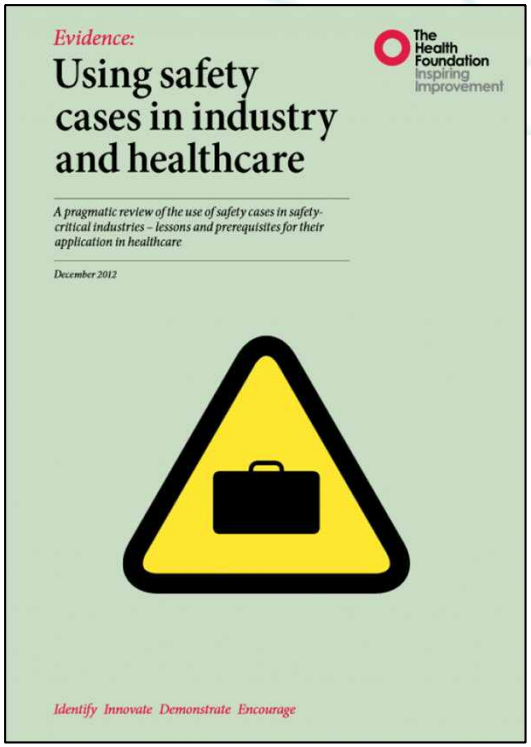
September 1998



Safety/Assurance Cases

Potential Benefits

- Promoting structured thinking about risk
- Fostering multidisciplinary communication about safety
- Integrating evidence sources
- Making the Implicit Explicit



From Safety Assurance to **Ethical** Assurance

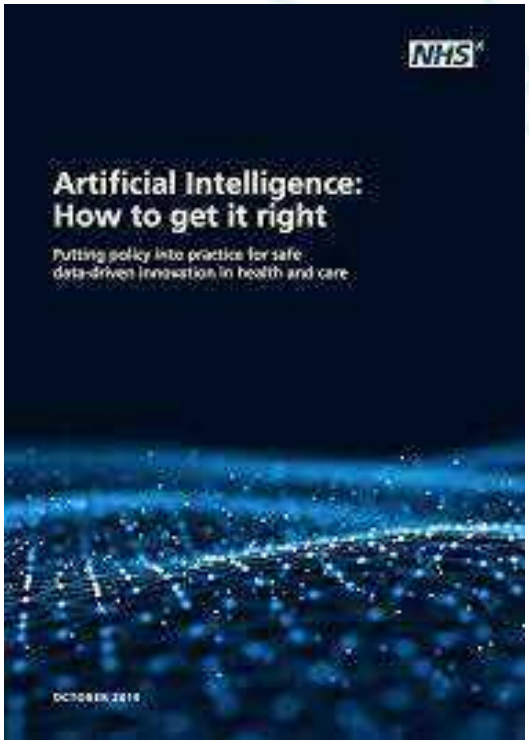
Ethical Assurance Cases: The What

Extending the assurance case methodology to achieve **justified confidence that a system will be *ethically acceptable* when used within a particular context**

Ethical Assurance Cases: The Why

Ethics is important

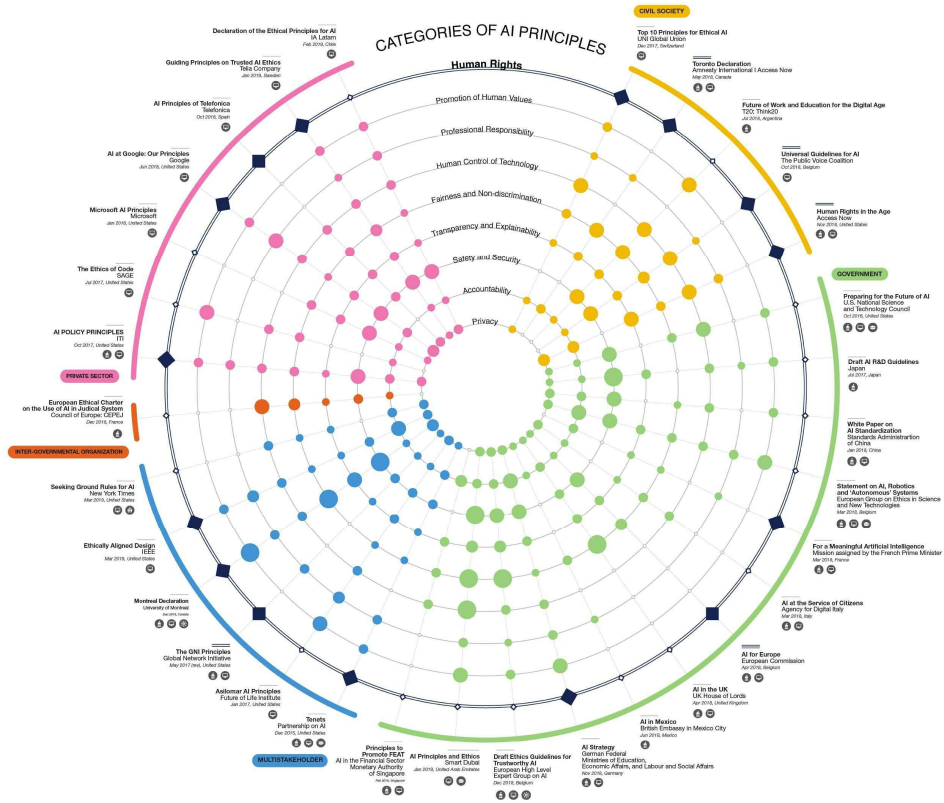
- For the technology to be positively transformative
- For the wellbeing of clinicians
- For the technology to gain traction
- For improved patient outcomes



Ethical Assurance Cases: The Why

Ethics covers a broad range of concerns

- Data Ethics (privacy, opacity, bias)
- Human agency and control
- Justice, fairness and responsibility

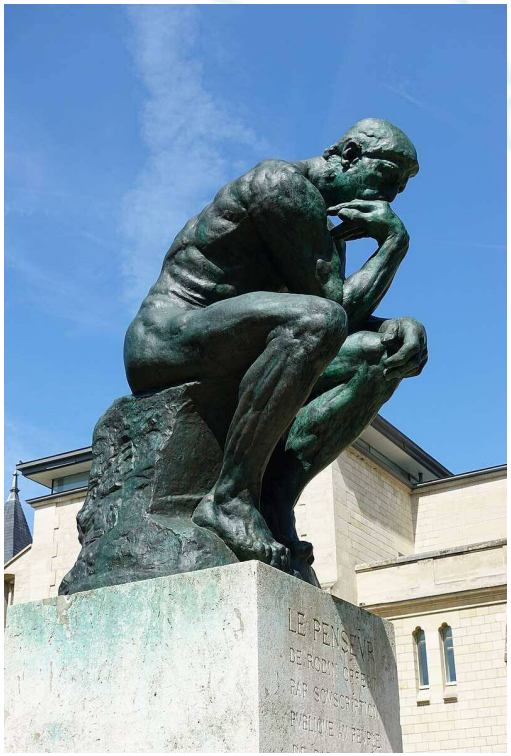


Berkman Klein Center for Internet and Society, Harvard University

Ethical Assurance Cases: The Why

Ethics is difficult

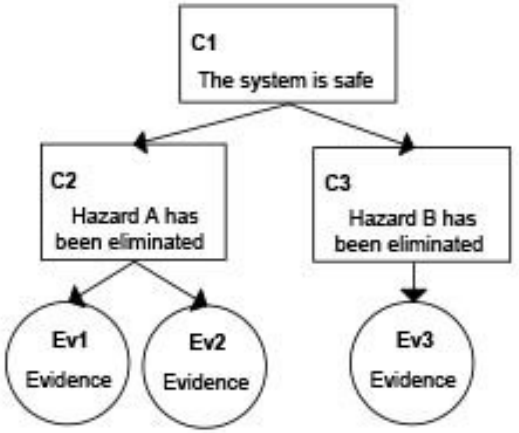
- No **single** right answer to many ethical questions
- Involves trade-offs
- Involves incommensurable values



Ethical Assurance Cases: The How

An emerging approach in the AI/AS ethics landscape

- Use the methodology to assure ethical properties beyond safety
- Use a graphical notation to present the argument



Ethical Assurance Cases: The How

Several emerging approaches in the AI/AS ethics landscape

- Assure a single ethical property vs. a suite of ethical properties
- ‘Bottom-up’: Use a participatory design approach to design the whole argument
- ‘Top-down’: Use a principles-based approach to design the argument (and participatory design to instantiate and validate individual assurance cases)

Fraunhofer

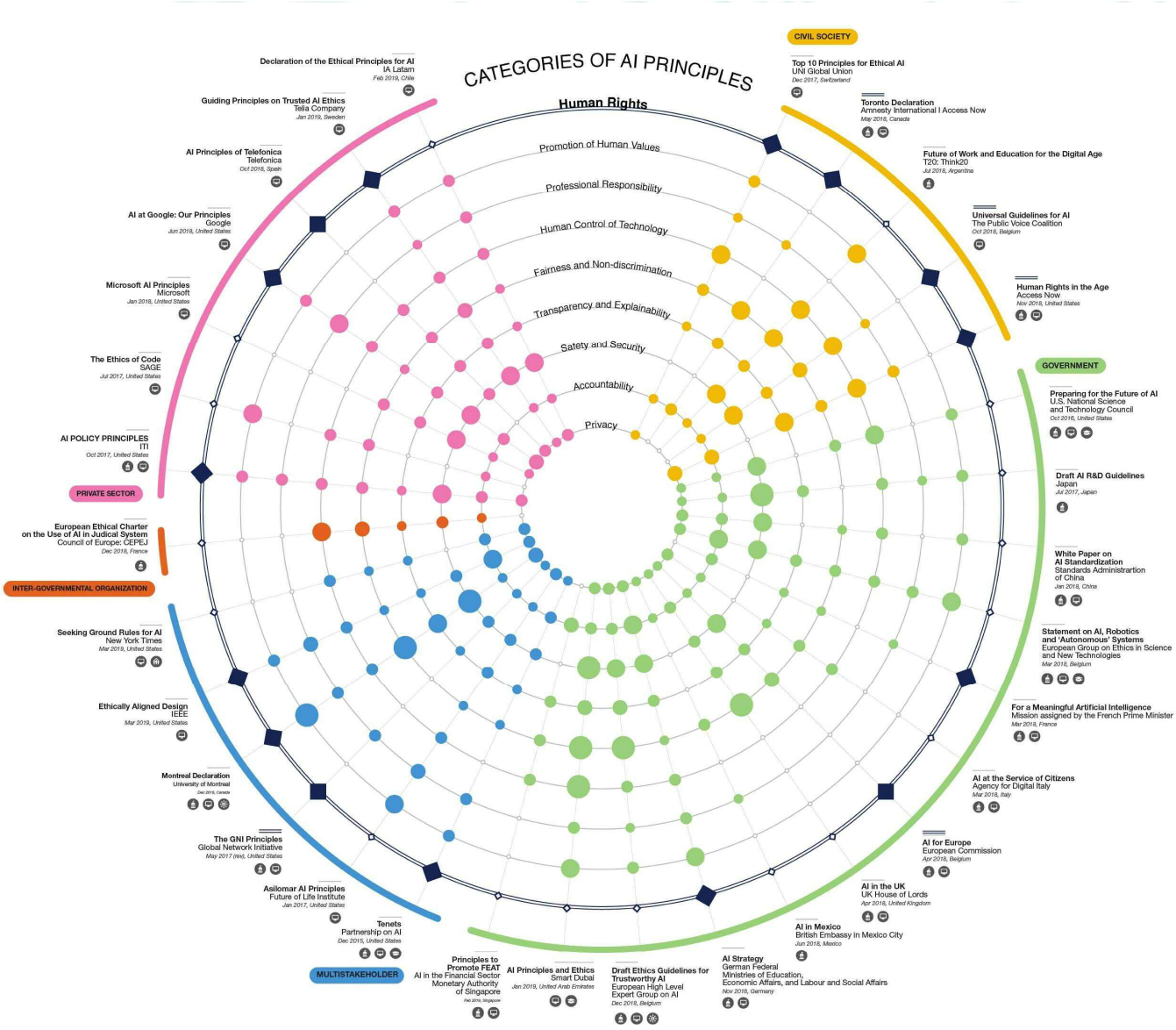
The Alan Turing Institute

ASSURING AUTONOMY INTERNATIONAL PROGRAMME

UNIVERSITY of York

Ethical principles

More than 80 major sets of ethical principles and ethics declarations published in the last few years of the 2010s – from government agencies and public bodies, NGOs, corporations, universities, and professional institutes.

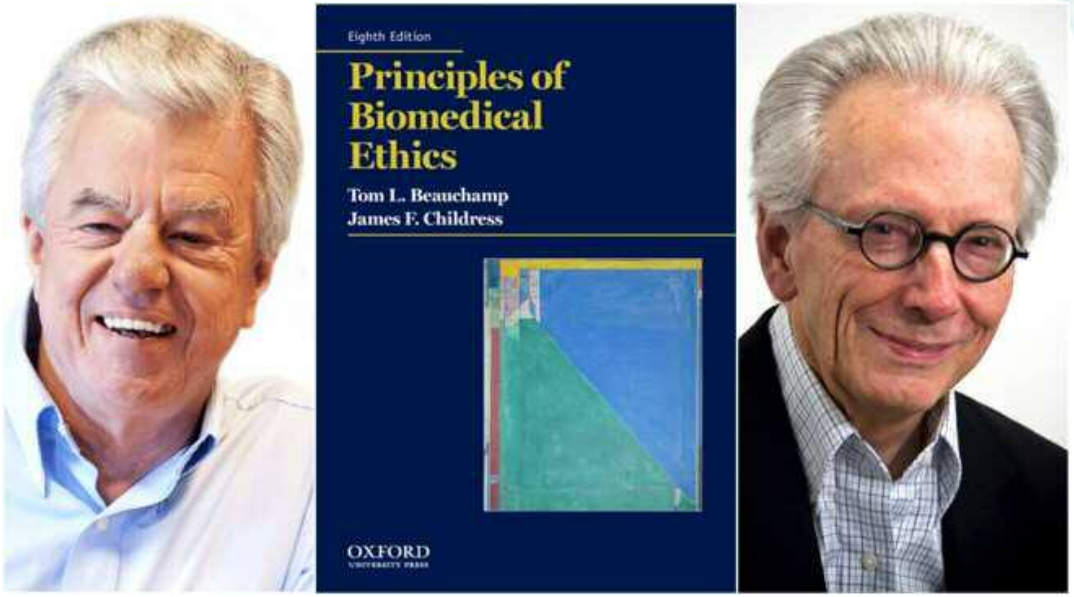


Source: Berkman Klein Center for Internet and Society, Harvard University

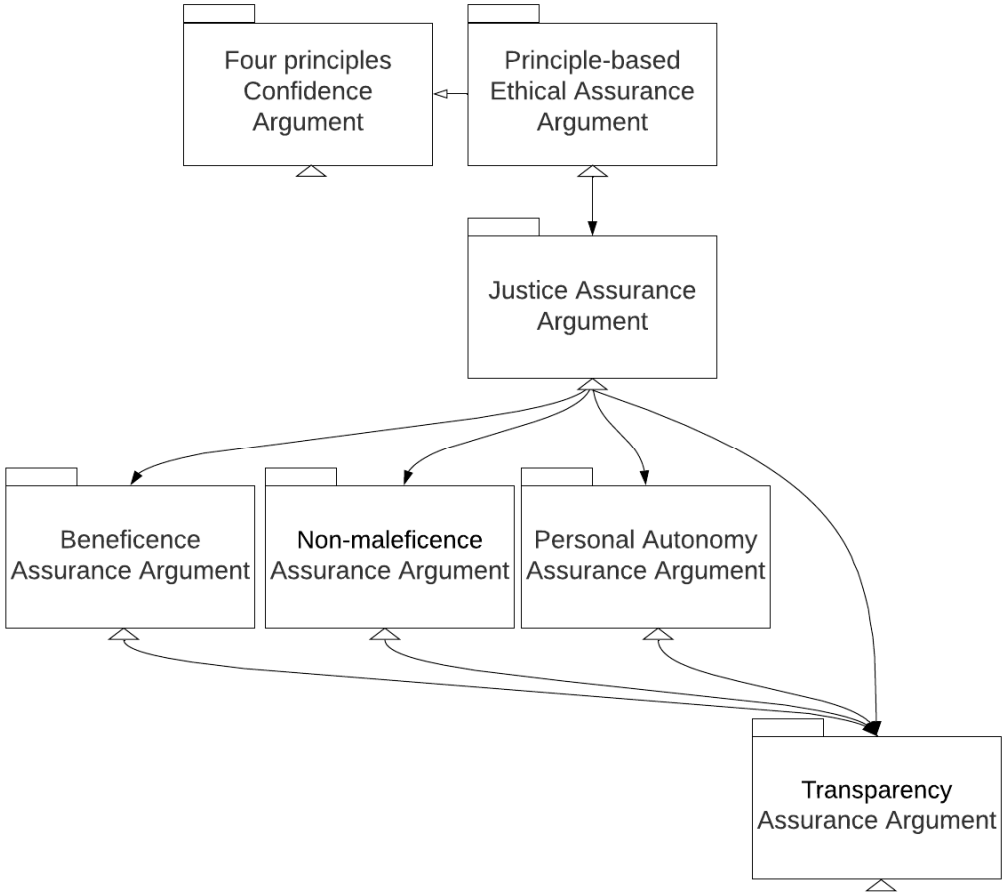
Four ethical principles

Striking overlap between these principles and the four classical principles of biomedical ethics:

- Non-maleficence
- Beneficence
- Respect for autonomy
- Justice

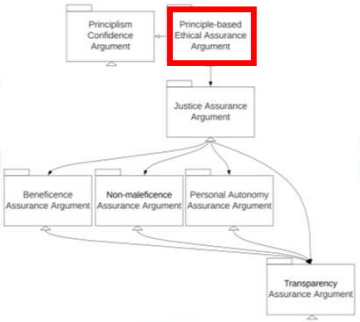


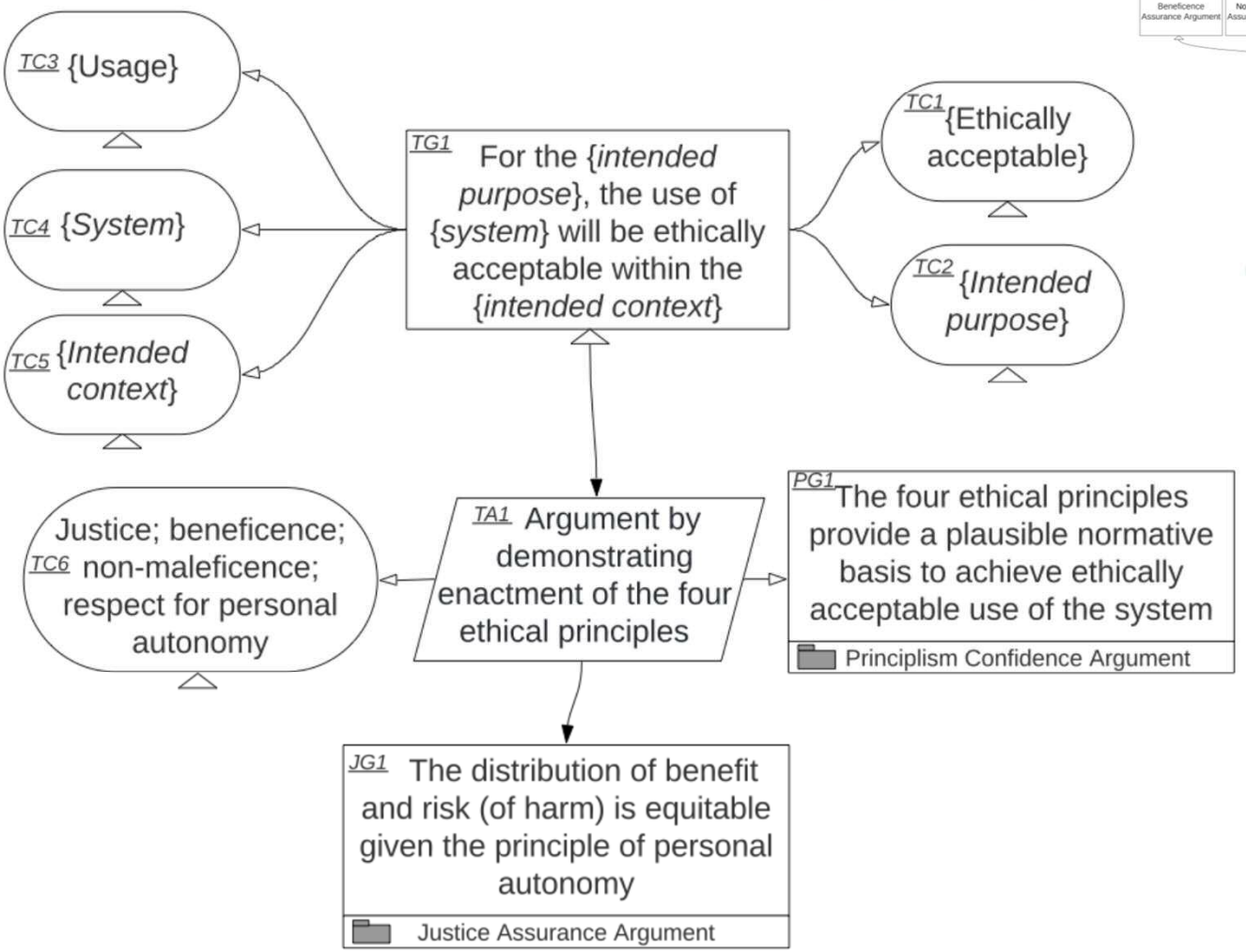
The Ethical Assurance Argument



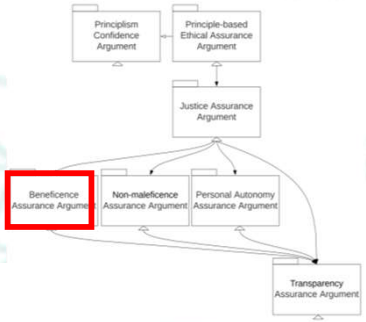
Ethical assurance argument

Based on the four ethical principles



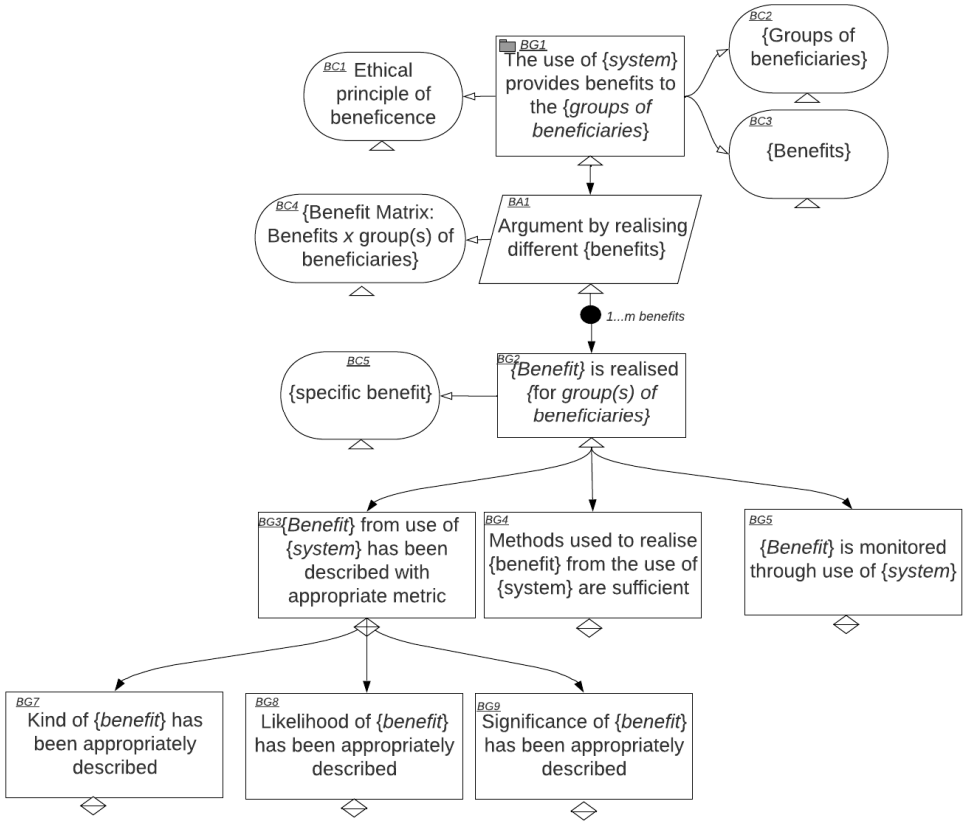


Beneficence argument



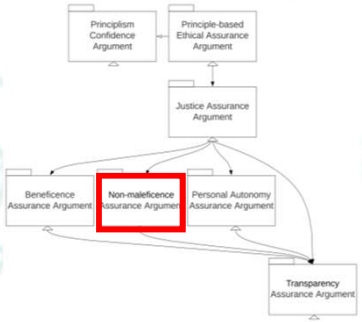
Do good

- What benefit does the proposed AI/AS promise for individuals, society or the environment?
- How are these benefits realized?
- Are they monitored over time?

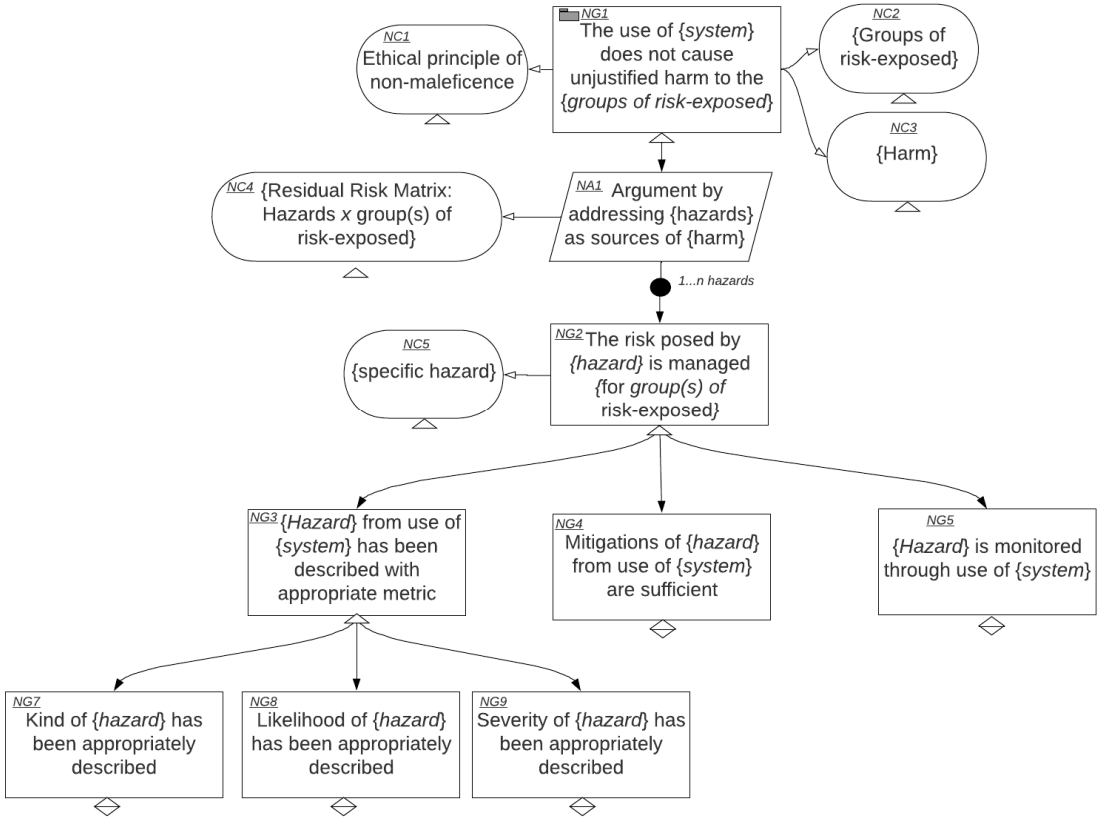


Non-maleficence argument

Do no (unjustified) harm

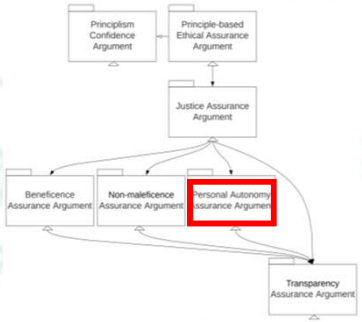


- What risks does the proposed AI/AS pose for individuals, society or the environment?
- How are these risks mitigated?
- Are they monitored over time?
- Range of harm from AI/AS extends beyond physical safety

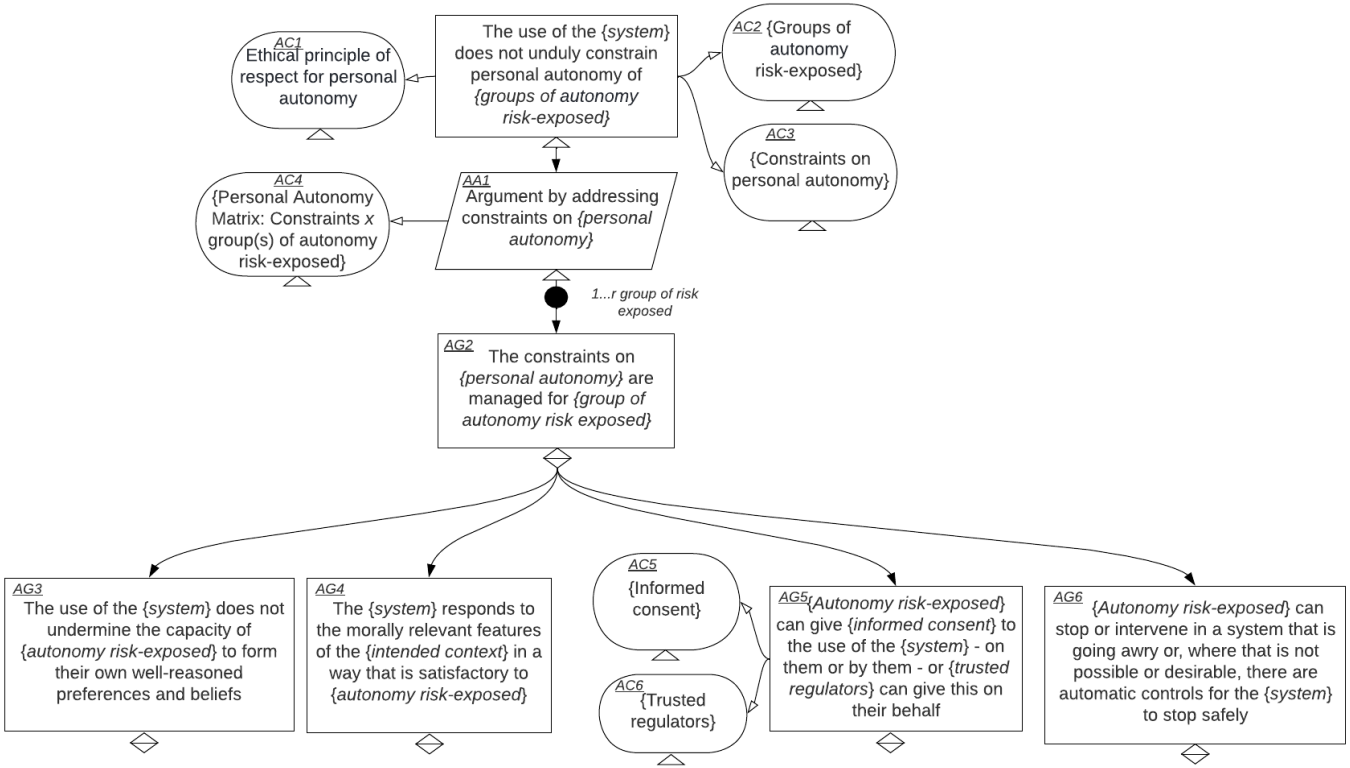


Personal autonomy argument

Respect people's autonomy

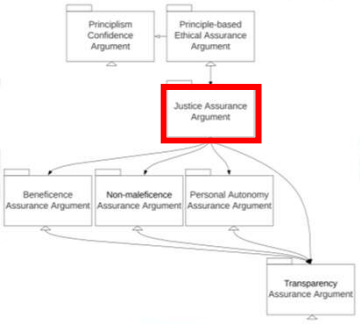


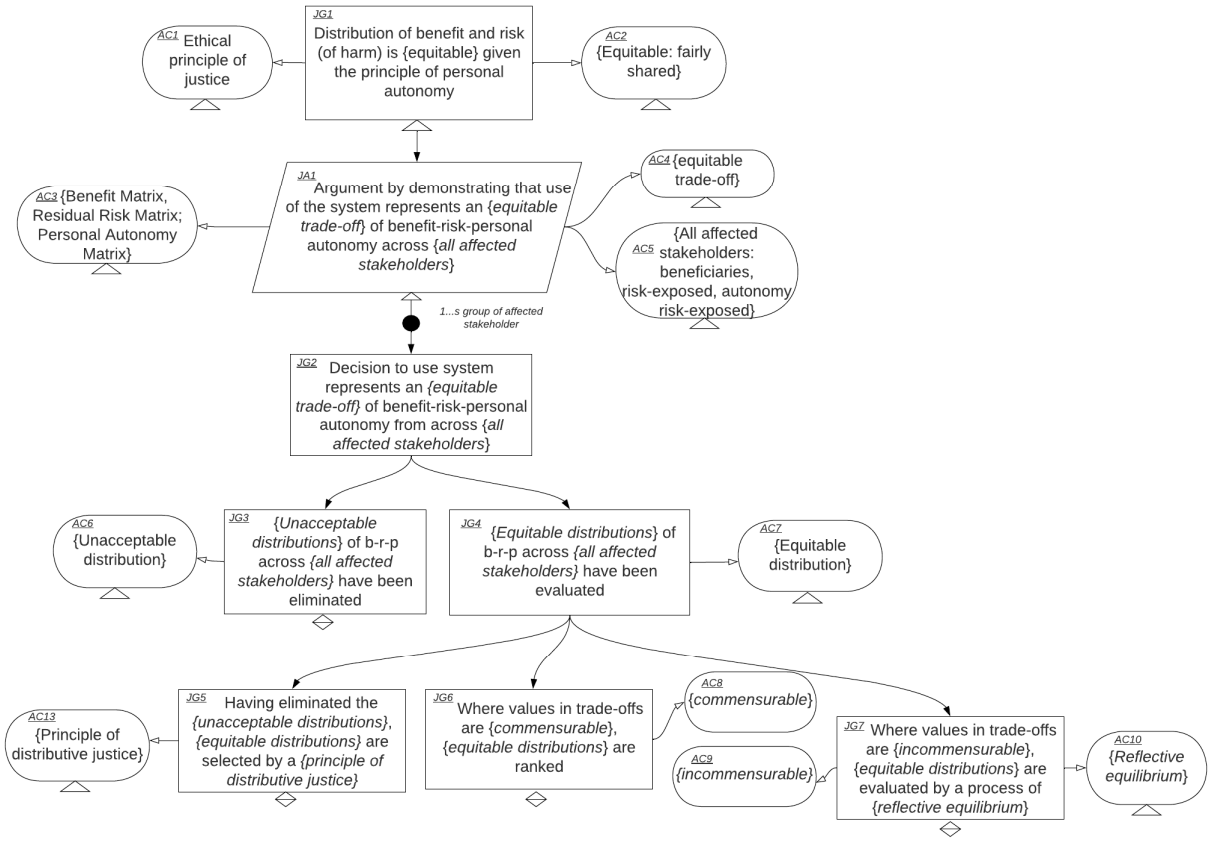
- Personal autonomy is central to moral agency and responsibility



Justice argument

- Where we reconcile trade-offs
- Need to consider the equitable distribution not just risk & benefit, but also agency (Hansson 2018)

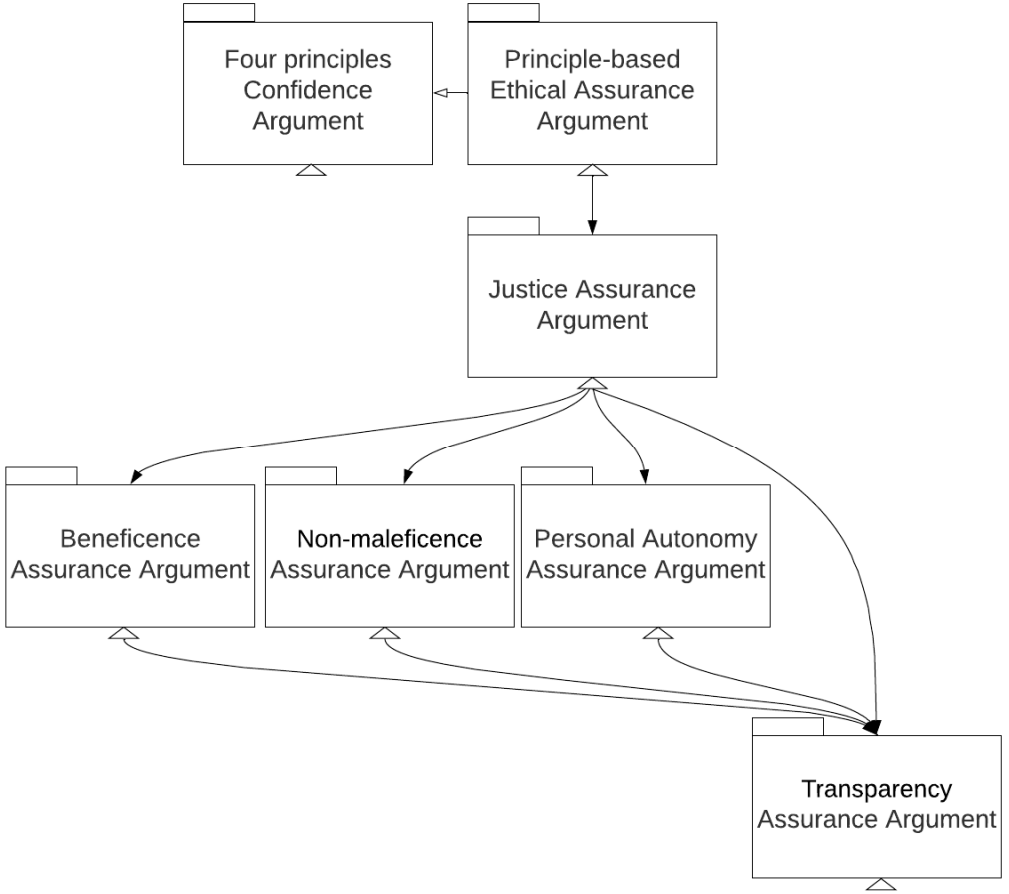




Transparency

An enabling condition

- An instrumental value rather than an end in itself
- What are the specific transparency requirements at each stage of the argument?



Deployment ethics

When would the decision to deploy an AI/AS be justified?

We need to deploy ethically defensible systems and learn from experience

Ethical assurance cases

A promising approach?

- Structural benefits
- Substantive benefits
- Communicative benefits

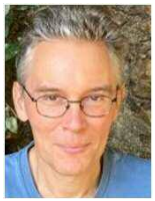
Extending the Ethical Assurance Methodology

New AS Responsibility Project

AR-TAS: Assuring Responsibility for Trustworthy Autonomous Systems





Lloyd's Register Foundation

Ufonia

NHS Digital

MIRA

SHEFFIELD ROBOTICS

WAYVE

JAGUAR

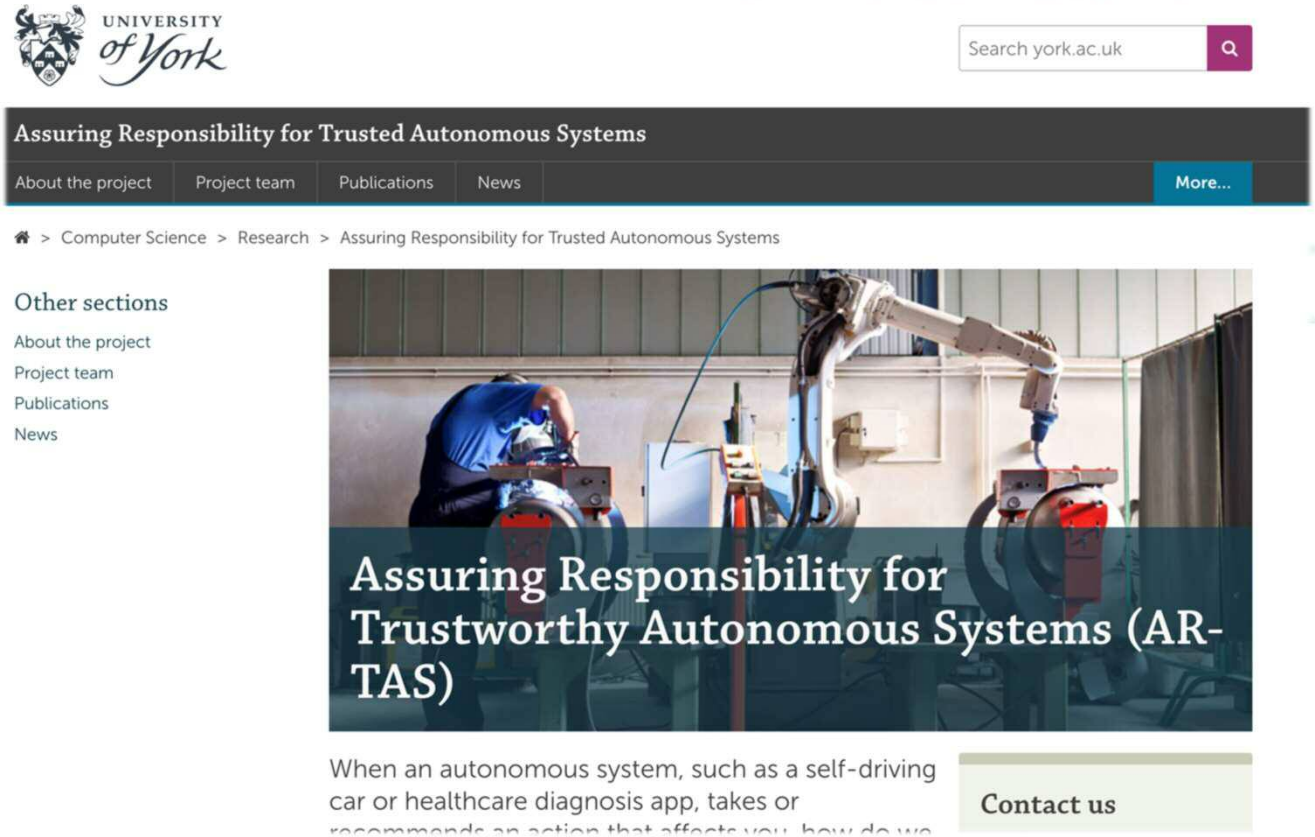
LAND ROVER

NASA

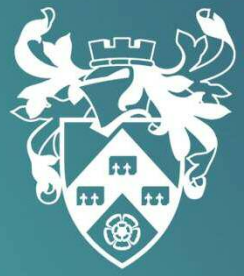
Project Reference: EP/W011239/1

Funded Period: Jan 22 - Jun 24

Funded Value: £703,615



https://www.cs.york.ac.uk/research/trusted-autonomous-systems/



UNIVERSITY of York

ASSURING AUTONOMY

INTERNATIONAL PROGRAMME