

# Beyond safety assurance: assuring the ethics of Al and autonomous systems



## 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

A **goal** presents a claim forming part of the argument.

A context presents a contextual artefact. This can be a reference to contextual information, or a statement.

argument has not been developed.Uninstantiated denotes that the attached element remains

Undeveloped element decorator indicates that a line of

A **strategy** describes the inference that exists between a goal and its supporting goal(s).

to be instantiated derives that the attached element remain to be instantiated, i.e. at some later stage the 'abstract' element needs to be replaced (instantiated) with a more concrete instance.

Public Decorator indicates that the element is publicly visible in one or more interfaces of the module and can

A **justification** presents a statement of rationale.

be referenced as an away element.

A solid ball is the symbol for **multiple** 

An **away goal** reference repeats a claim presented in another argument module (named below)

A solid diamond is the symbol for Choice.

instantiations.

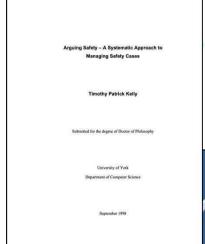
Argument module name

A solution presents a reference to an evidence item.

A **module** presents a reference to a module containing an argument

**SupportedBy** allows support relationships between elements to be documented.

InContextOf declares a contextual relationship.



Goal Structuring Notation Community Standard Version 3

> The Assurance Case Working Group (ACWG)

> > SCSC-141C

## 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

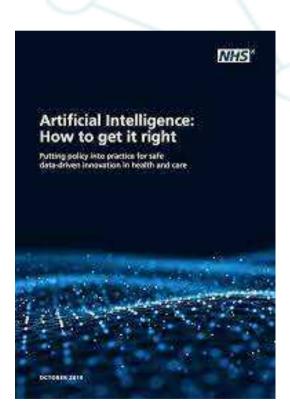
### Ethical Assurance Cases: The What

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**

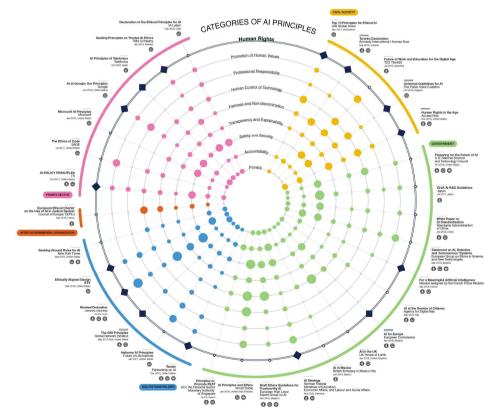
- For the technology to be positively transformative
- For the wellbeing of clinicians
- For the technology to gain traction
- For improved patient outcomes



## 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

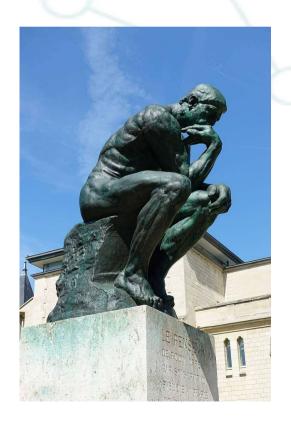


Berkman Klein Center for Internet and Society, Harvard University

## Ethical Assurance Cases: The Why

#### **Ethics is difficult**

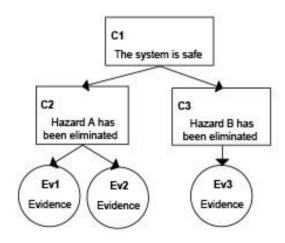
- No single right answer to many ethical questions
- Involves trade-offs
- Involves incommensurable values



### 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)



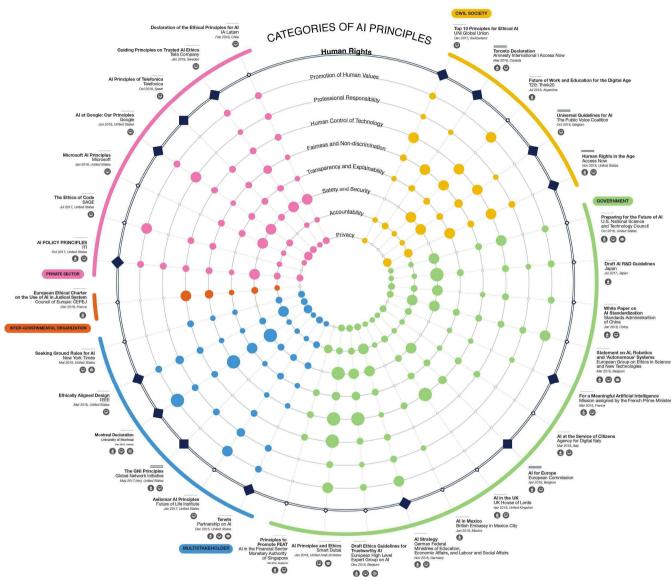






# Ethical principles

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.

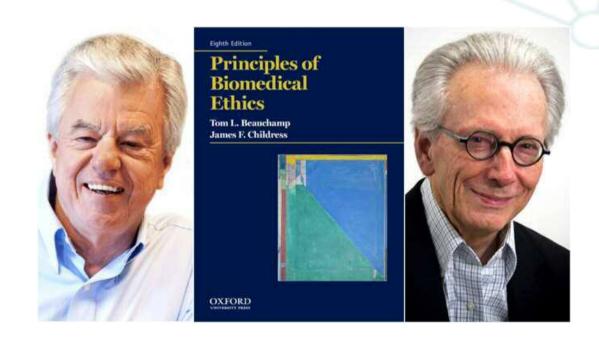


Source: Berkman Klein Center for Internet and Society, Harvard University

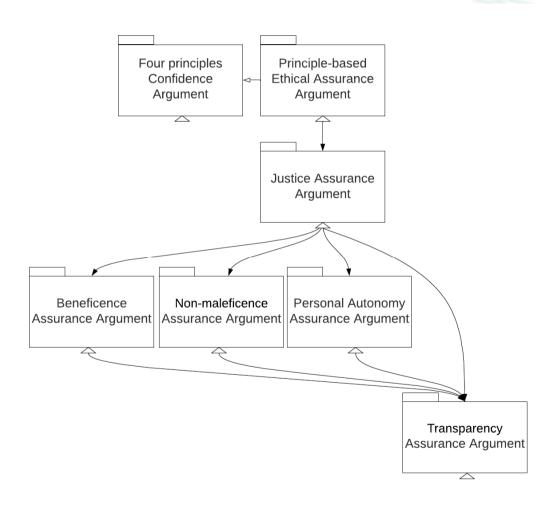
## Four ethical principles

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

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

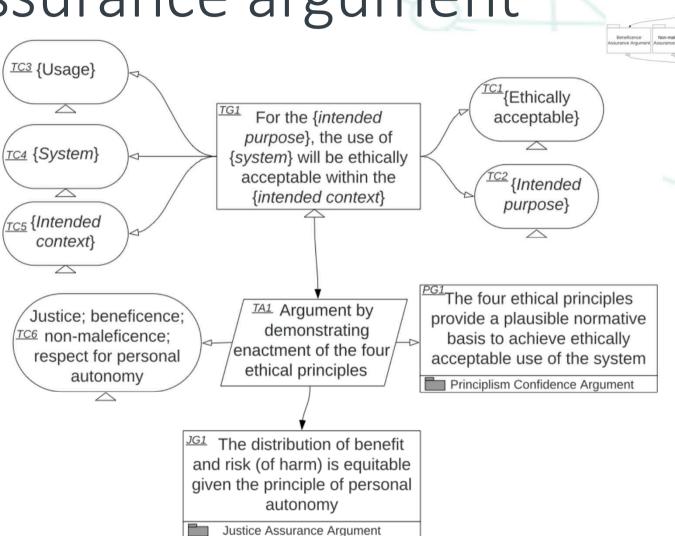


## The Ethical Assurance Argument



## Ethical assurance argument

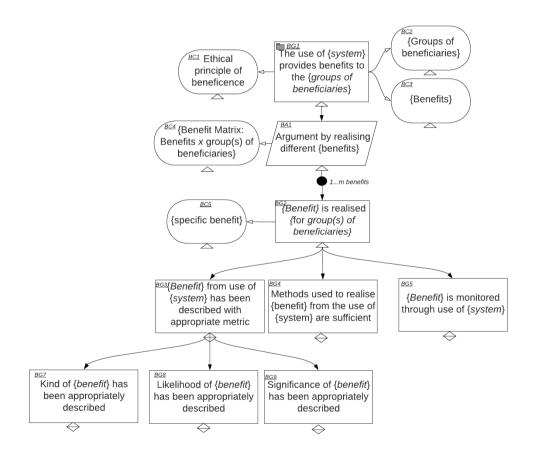
Based on the four ethical principles



### Beneficence argument

### Do good

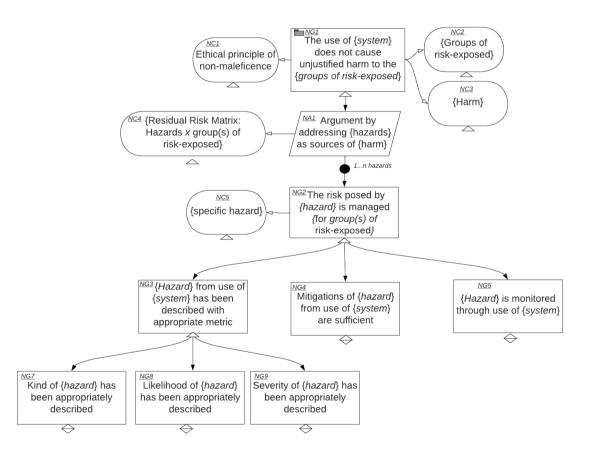
- What benefit does the proposed AI/AS promise for individuals, society or the environment?
- How are these benefits realized?
- Are they monitored over time?



# Non-maleficence argument

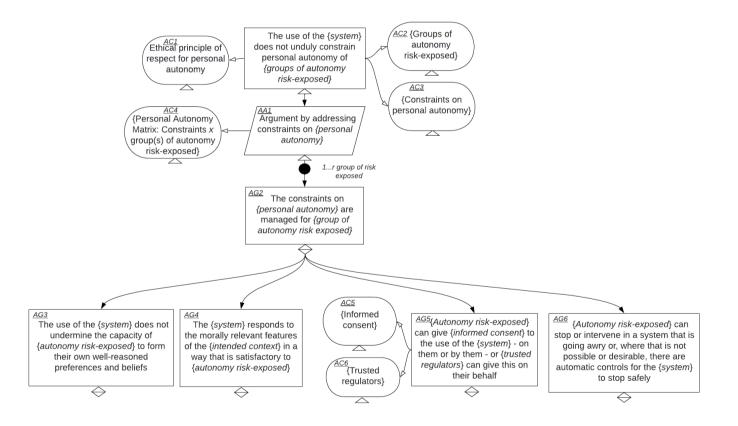
Do no (unjustified) harm

- 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



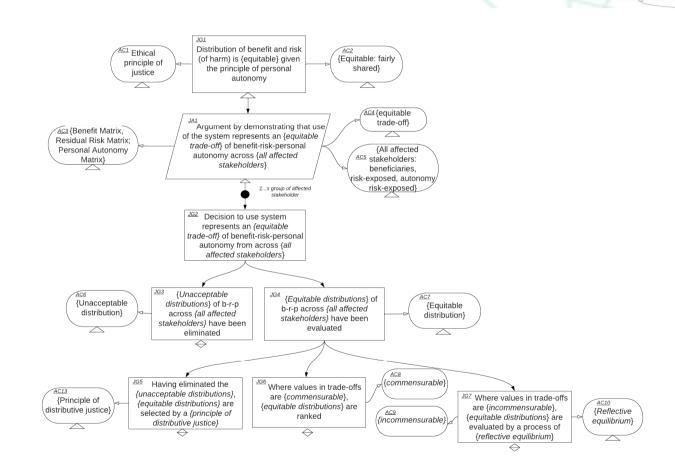
# Personal autonomy argument Respect people's autonomy

 Personal autonomy is central to moral agency and responsibility



### Justice argument

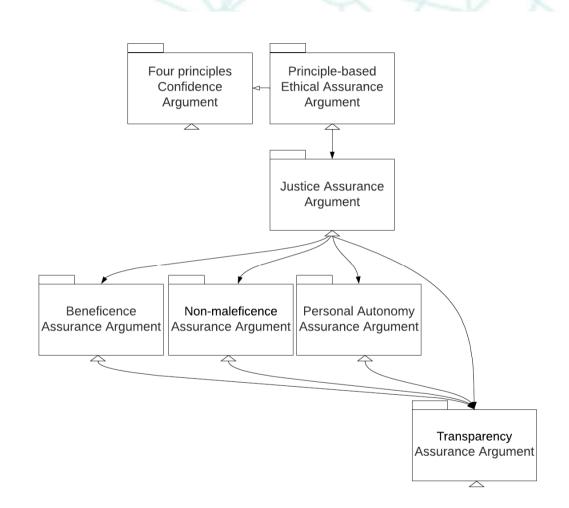
- Where we reconcile trade-offs
- Need to consider the equitable distribution not just risk & benefit, but also agency (Hansson 2018)



### 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



















**Ufonia** 











Search york.ac.uk

#### Assuring Responsibility for Trusted Autonomous Systems

Publications

\* > Computer Science > Research > Assuring Responsibility for Trusted Autonomous Systems

#### Other sections

About the project Project team **Publications** 

News

Assuring Responsibility for Trustworthy Autonomous Systems (AR-TAS)

When an autonomous system, such as a self-driving car or healthcare diagnosis app, takes or

recommends an action that affects you how do we

Contact us

Project Reference: Funded Period: EP/W011239/1

Jan 22 - Jun 24

**Funded Value:** £703,615

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



### **ASSURING AUTONOMY**

**INTERNATIONAL PROGRAMME**