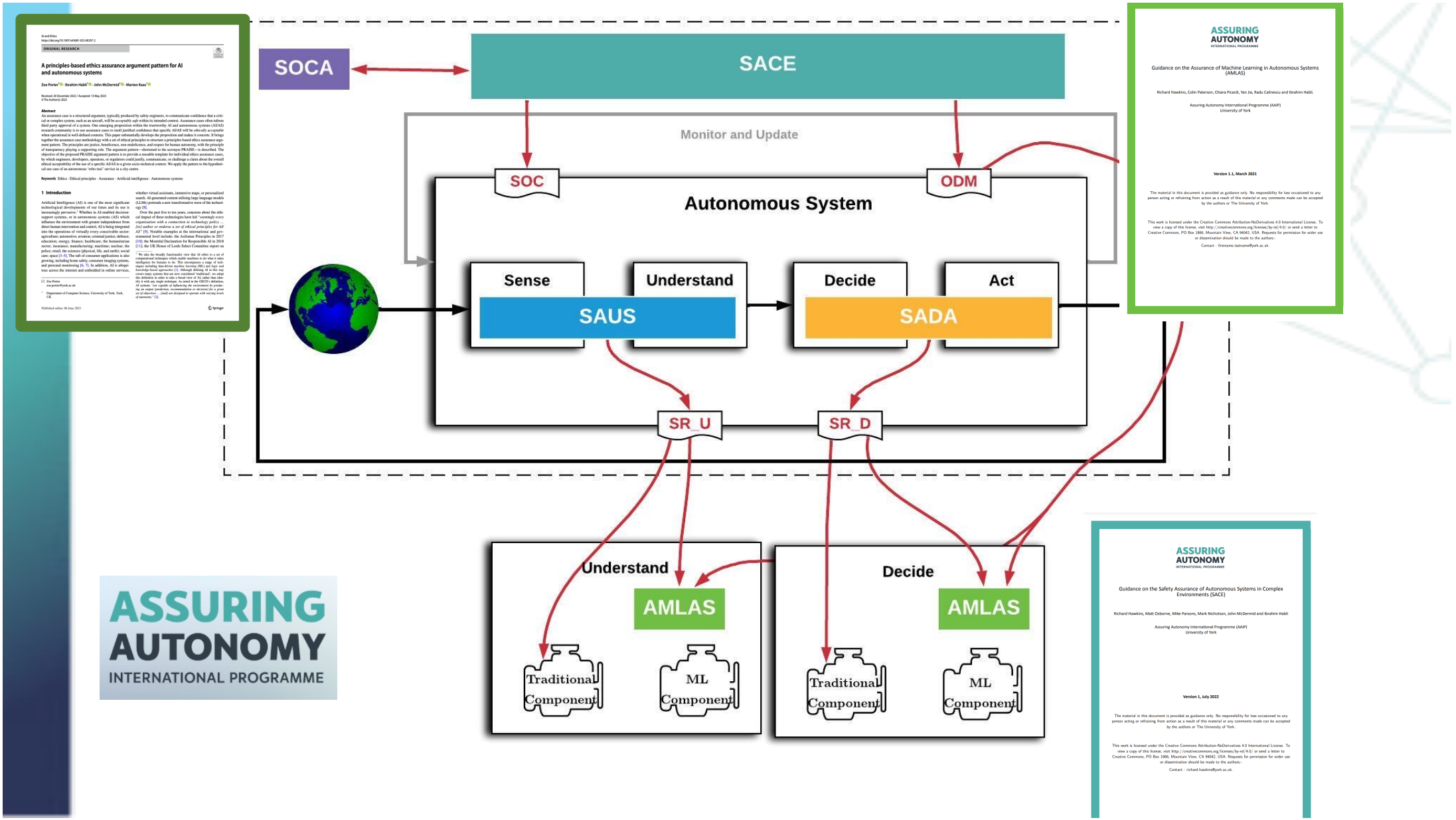
UNIVERSITY of York

ASSURING AUTONOMY
INTERNATIONAL PROGRAMME

A Principles-based Ethical Assurance Argument for AI & Autonomous Systems

Professor Ibrahim Habli

A Whole System Approach to Assurance



AI and Ethics
https://doi.org/10.1007/s43681-023-00297-2

ORIGINAL RESEARCH

A principles-based ethics assurance argument pattern for AI and autonomous systems

Zoe Porter[1] · Ibrahim Habli[1] · John McDermid[1] · Marten Kaas[1]

Received: 20 December 2022 / Accepted: 13 May 2023

Abstract
An assurance case is a structured argument, typically produced by safety engineers, to communicate confidence that a critical or complex system, such as an aircraft, will be *acceptably safe* within its intended context. Assurance cases often inform third party approval of a system. One emerging proposition within the trustworthy AI and autonomous systems (AI/AS) research community is to use assurance cases to instil justified confidence that specific AI/AS will be ethically *acceptable* when operational in well-defined contexts. This paper substantially develops the proposition and makes it concrete. It brings together the assurance case methodology with a set of ethical principles to structure a principles-based ethics assurance argument pattern. The principles are justice, beneficence, non-maleficence, and respect for human autonomy, with the principle of transparency playing a supporting role. The argument pattern—shortened to the acronym PRAISE—is described. The objective of the proposed PRAISE argument pattern is to provide a reusable template for individual ethics assurance cases, by which engineers, developers, operators, or regulators could justify, communicate, or challenge a claim about the overall ethical acceptability of the use of a specific AI/AS in a given socio-technical context. We apply the pattern to the hypothetical use case of an autonomous 'robo-taxi' service in a city centre.

Keywords Ethics · Ethical principles · Assurance · Artificial intelligence · Autonomous systems

1 Introduction

Artificial Intelligence (AI) is one of the most significant technological developments of our times and its use is increasingly pervasive.[1] Whether in AI-enabled decision-support systems, or in autonomous systems (AS) which influence the environment with greater independence from direct human intervention and control, AI is being integrated into the operations of virtually every conceivable sector: agriculture; automotive; aviation; criminal justice; defence; education; energy; finance; healthcare; the humanitarian sector; insurance; manufacturing; maritime; nuclear; the police; retail; the sciences (physical, life, and earth); social care; space [3–5]. The raft of consumer applications is also growing, including home safety, consumer imaging systems, and personal monitoring [6, 7]. In addition, AI is ubiquitous across the internet and embedded in online services, whether virtual assistants, immersive maps, or personalised search. AI-generated content utilising large language models (LLMs) portends a new transformative wave of the technology [8].

Over the past five to ten years, concerns about the ethical impact of these technologies have led "*seemingly every organisation with a connection to technology policy … [to] author or endorse a set of ethical principles for AI*" [9]. Notable examples at the international and governmental level include: the Asilomar Principles in 2017 [10]; the Montréal Declaration for Responsible AI in 2018 [11]; the UK House of Lords Select Committee report on

[1] We take the broadly functionalist view that AI refers to a set of computational techniques which enable machines to do what it takes intelligence for humans to do. This encompasses a range of techniques including data-driven *machine learning* (ML) and *logic and knowledge-based approaches* [2]. To ensure that the PRAISE argument covers many systems that are now considered 'traditional', we adopt this definition in order to take a broad view of AI, rather than identify it with any single technique. As noted in the OECD's definition, AI systems "*are capable of influencing the environment by producing an output (predictions, recommendations or decisions) for a given set of objectives … [and] are designed to operate with varying levels of autonomy.*" [2].

✉ Zoe Porter
zoe.porter@york.ac.uk

[1] Department of Computer Science, University of York, York, UK

Published online: 06 June 2023

ASSURING AUTONOMY
INTERNATIONAL PROGRAMME

Guidance on the Assurance of Machine Learning in Autonomous Systems (AMLAS)

Richard Hawkins, Colin Paterson, Chiara Picardi, Yan Jia, Radu Calinescu and Ibrahim Habli.

Assuring Autonomy International Programme (AAIP)
University of York

Version 1.1, March 2021

ASSURING AUTONOMY
INTERNATIONAL PROGRAMME

Guidance on the Safety Assurance of Autonomous Systems in Complex Environments (SACE)

Richard Hawkins, Matt Osborne, Mike Parsons, Mark Nicholson, John McDermid and Ibrahim Habli

Assuring Autonomy International Programme (AAIP)
University of York

Version 1, July 2022

ASSURING AUTONOMY
INTERNATIONAL PROGRAMME

AI and Ethics
https://doi.org/10.1007/s43681-023-00297-2

ORIGINAL RESEARCH

A principles-based ethics assurance argument pattern for AI and autonomous systems

Zoe Porter[1] · Ibrahim Habli[1] · John McDermid[1] · Marten Kaas[1]

Received: 20 December 2022 / Accepted: 13 May 2023

Abstract
An assurance case is a structured argument, typically produced by safety engineers, to communicate confidence that a critical or complex system, such as an aircraft, will be *acceptably safe* within its intended context. Assurance cases often inform third party approval of a system. One emerging proposition within the trustworthy AI and autonomous systems (AI/AS) research community is to use assurance cases to instil justified confidence that specific AI/AS will be *ethically acceptable* when operational in well-defined contexts. This paper substantially develops the proposition and makes it concrete. It brings together the assurance case methodology with a set of ethical principles to structure a principles-based ethics assurance argument pattern. The principles are justice, beneficence, non-maleficence, and respect for human autonomy, with the principle of transparency playing a supporting role. The argument pattern—shortened to the acronym PRAISE—is described. The objective of the proposed PRAISE argument pattern is to provide a reusable template for individual ethics assurance cases, by which engineers, developers, operators, or regulators could justify, communicate, or challenge a claim about the overall ethical acceptability of the use of a specific AI/AS in a given socio-technical context. We apply the pattern to the hypothetical use case of an autonomous 'robo-taxi' service in a city centre.

Keywords Ethics · Ethical principles · Assurance · Artificial intelligence · Autonomous systems

1 Introduction

Artificial Intelligence (AI) is one of the most significant technological developments of our times and its use is increasingly pervasive.[1] Whether in AI-enabled decision-support systems, or in autonomous systems (AS) which influence the environment with greater independence from direct human intervention and control, AI is being integrated into the operations of virtually every conceivable sector: agriculture; automotive; aviation; criminal justice; defence; education; energy; finance; healthcare; the humanitarian sector; insurance; manufacturing; maritime; nuclear; the police; retail; the sciences (physical, life, and earth); social care; space [3–5]. The raft of consumer applications is also growing, including home safety, consumer imaging systems, and personal monitoring [6, 7]. In addition, AI is ubiquitous across the internet and embedded in online services, whether virtual assistants, immersive maps, or personalised search. AI-generated content utilising large language models (LLMs) portends a new transformative wave of the technology [8].

Over the past five to ten years, concerns about the ethical impact of these technologies have led *"seemingly every organisation with a connection to technology policy … [to] author or endorse a set of ethical principles for AI/AS"* [9]. Notable examples at the international and governmental level include: the Asilomar Principles in 2017 [10]; the Montréal Declaration for Responsible AI in 2018 [11]; the UK House of Lords Select Committee report on

✉ Zoe Porter
zoe.porter@york.ac.uk

[1] Department of Computer Science, University of York, York, UK

[1] We take the broadly functionalist view that AI refers to a set of computational techniques which enable machines to do what it takes intelligence for humans to do. This encompasses a range of techniques including data-driven *machine learning* (ML) and *logic and knowledge-based approaches* [1]. Although defining AI in this way covers many systems that are now considered 'traditional', we adopt this definition in order to take a broad view of AI, rather than identify it with any single technique. As noted in the OECD's definition, AI systems *"are capable of influencing the environment by producing an output (prediction, recommendation or decision) for a given set of objectives … [and] are designed to operate with varying levels of autonomy."* [2].

Published online: 06 June 2023

ASSURING AUTONOMY
INTERNATIONAL PROGRAMME

Guidance on the Safety Assurance of Autonomous Systems in Complex Environments (SACE)

Richard Hawkins, Matt Osborne, Mike Parsons, Mark Nicholson, John McDermid and Ibrahim Habli

Assuring Autonomy International Programme (AAIP)
University of York

Version 1, July 2022

Contact : richard.hawkins@york.ac.uk.

ASSURING AUTONOMY
INTERNATIONAL PROGRAMME

Guidance on the Assurance of Machine Learning in Autonomous Systems (AMLAS)

Richard Hawkins, Colin Paterson, Chiara Picardi, Yan Jia, Radu Calinescu and Ibrahim Habli.

Assuring Autonomy International Programme (AAIP)
University of York

Version 1.1, March 2021

Contact : firstname.lastname@york.ac.uk.

AI and Ethics
https://doi.org/10.1007/s43681-023-00297-2

ORIGINAL RESEARCH

A principles-based ethics assurance argument pattern for AI and autonomous systems

Zoe Porter[1] · Ibrahim Habli[1] · John McDermid[1] · Marten Kaas[1]

Received: 20 December 2022 / Accepted: 13 May 2023

Abstract
An assurance case is a structured argument, typically produced by safety engineers, to communicate confidence that a critical or complex system, such as an aircraft, will be *acceptably safe* within its intended context. Assurance cases often inform third party approval of a system. One emerging proposition within the trustworthy AI and autonomous systems (AI/AS) research community is to use assurance cases to instil justified confidence that specific AI/AS will be *ethically acceptable* when operational in well-defined contexts. This paper substantially develops the proposition and makes it concrete. It brings together the assurance case methodology with a set of ethical principles to structure a principles-based ethics assurance argument pattern. The principles are justice, beneficence, non-maleficence, and respect for human autonomy, with the principle of transparency playing a supporting role. The argument pattern—shortened to the acronym PRAISE—is described. The objective of the proposed PRAISE argument pattern is to provide a reusable template for individual ethics assurance cases, by which engineers, developers, operators, or regulators could justify, communicate, or challenge a claim about the overall ethical acceptability of the use of a specific AI/AS in a given socio-technical context. We apply the pattern to the hypothetical use case of an autonomous 'robo-taxi' service in a city centre.

Keywords Ethics · Ethical principles · Assurance · Artificial intelligence · Autonomous systems

1 Introduction

Artificial Intelligence (AI) is one of the most significant technological developments of our times and its use is increasingly pervasive.[1] Whether in AI-enabled decision-support systems, or in autonomous systems (AS) which influence the environment with greater independence from direct human intervention and control, AI is being integrated into the operations of virtually every conceivable sector: agriculture; automotive; aviation; criminal justice; defence; education; energy; finance; healthcare; the humanitarian sector; insurance; manufacturing; maritime; nuclear; the police; retail; the sciences (physical, life, and earth); social care; space [3–5]. The raft of consumer applications is also growing, including home safety, consumer imaging systems, and personal monitoring [6, 7]. In addition, AI is ubiquitous across the internet and embedded in online services, whether virtual assistants, immersive maps, or personalised search. AI-generated content utilising large language models (LLMs) portends a new transformative wave of the technology [8].

Over the past five to ten years, concerns about the ethical impact of these technologies have led *"seemingly every organisation with a connection to technology policy … [to] author or endorse a set of ethical principles for AI/AS"* [9]. Notable examples at the international and governmental level include: the Asilomar Principles in 2017 [10]; the Montréal Declaration for Responsible AI in 2018 [11]; the UK House of Lords Select Committee report on

✉ Zoe Porter
zoe.porter@york.ac.uk

[1] Department of Computer Science, University of York, York, UK

[1] We take the broadly functionalist view that AI refers to a set of computational techniques which enable machines to do what it takes intelligence for humans to do. This encompasses a range of techniques including data-driven *machine learning* (ML) and *logic and knowledge-based approaches* [1]. Although defining AI in this way covers many systems that are now considered 'traditional', we adopt this definition in order to take a broad view of AI, rather than identify it with any single technique. As noted in the OECD's definition, AI systems *"are capable of influencing the environment by producing an output (prediction, recommendation or decision) for a given set of objectives … [and] are designed to operate with varying levels of autonomy."* [2].

Published online: 06 June 2023

Safety/Assurance Cases

- Paradigm shift in many domains
  - Shift from a prescribed process to a product-oriented assurance
  - Shift from a tick-box to argument-based

- Different drivers:
  - Accidents
    - Piper Alpha, 1988
  - Incidents and recalls
    - FDA, 2010
  - Complexity
    - Automotive, 2011
  - Greater complexity through AI
    - Autonomous driving, 2015

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

Back to the basics

Safety ~~and~~ is ethics

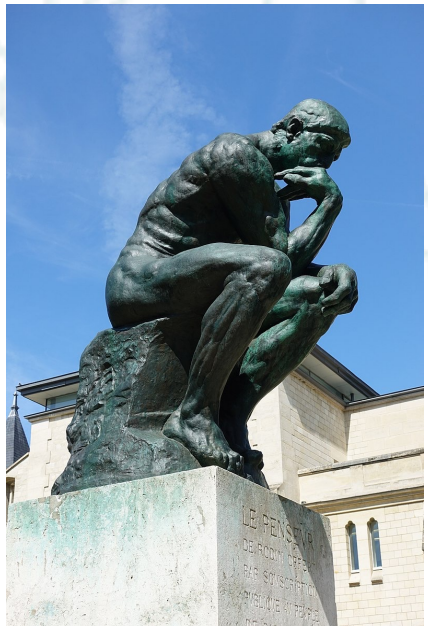
Many safety concerns are ethical concerns

- Harm-avoidance and proportionate risk are classic safety concerns, but they are also ethical concerns
- Just culture and human control/autonomy are ethical concerns which can have an impact on safety

Ethical Assurance

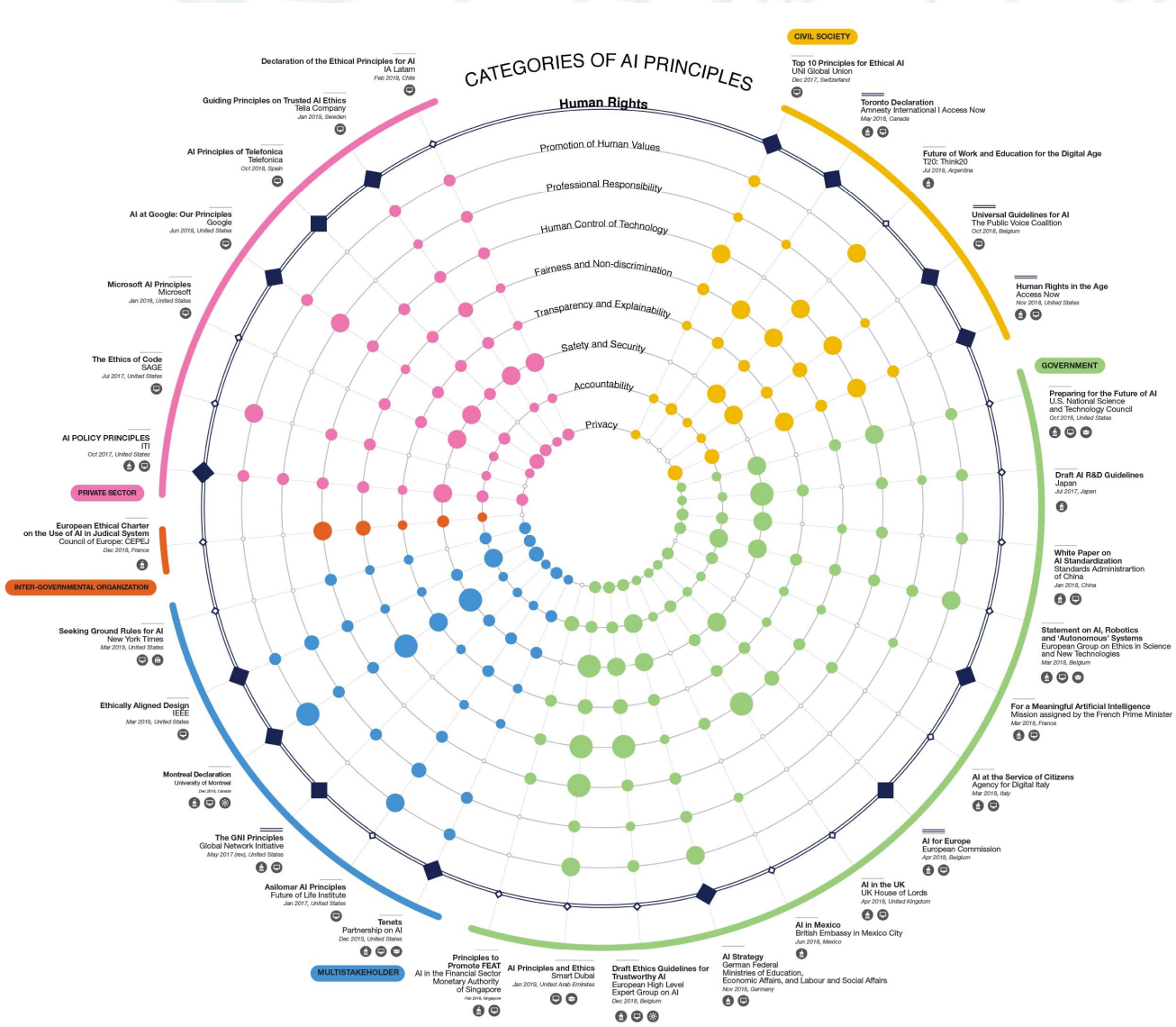
A definition

AI/AS will be ethically acceptable if affected stakeholders could not reasonably reject the decision to deploy it



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.

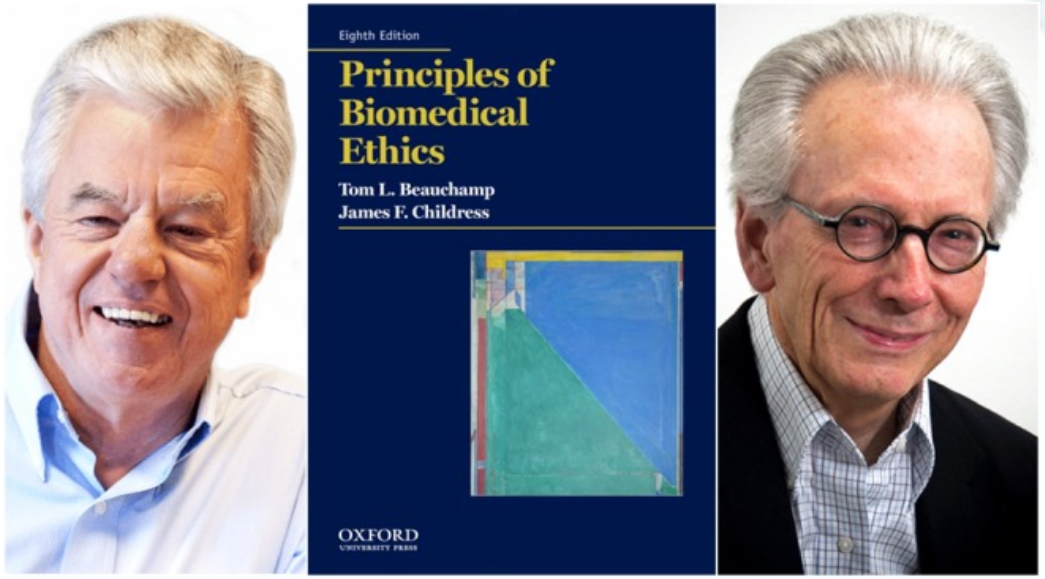


Source: Berkman Klein Center for Internet and Society, Harvard University

Four ethical principles

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

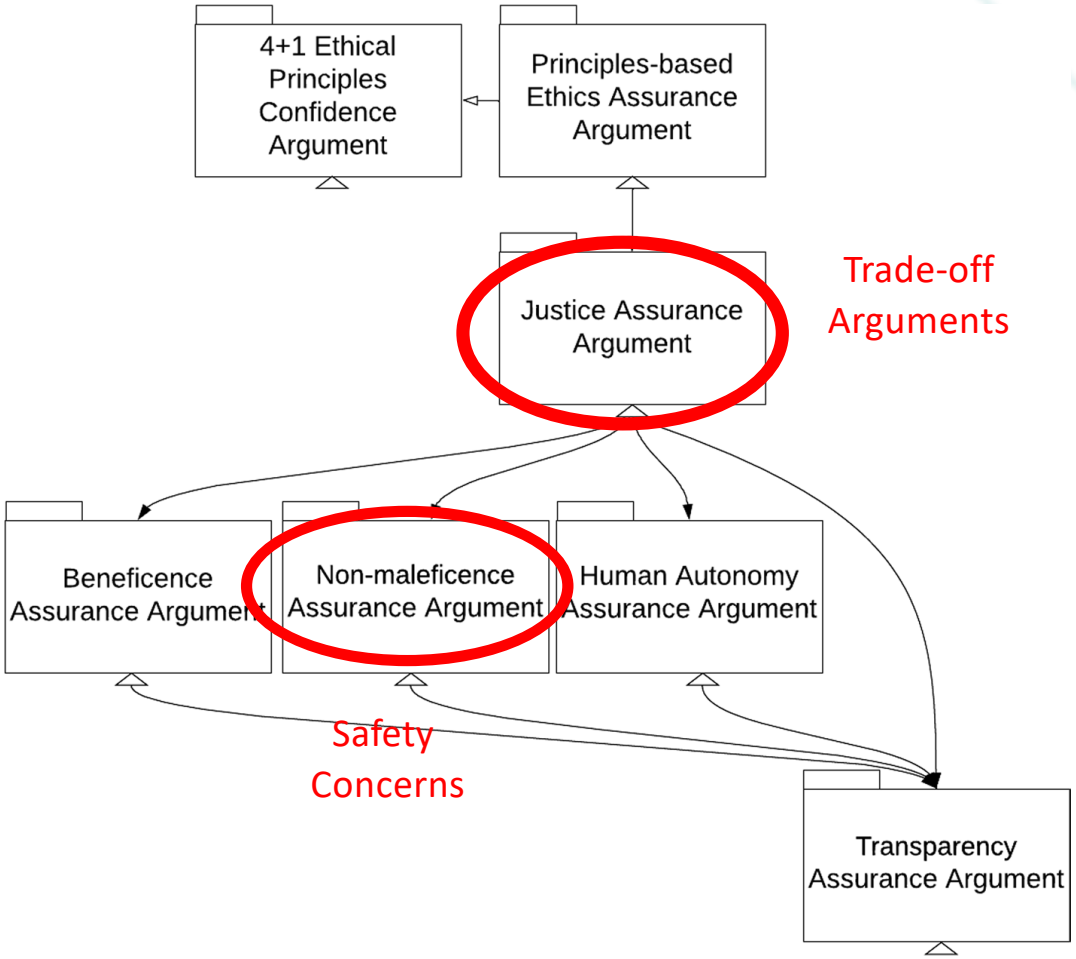
- Non-maleficence
- Beneficence
- Respect for autonomy
- Justice



Four ethical principles

- **Justice**: the distribution of benefits and risks from use of the system should be equitable across affected stakeholders
- **Beneficence**: the use of the system should benefit affected stakeholders
- **Non-maleficence**: the use of the system should not cause unjustified harm to affected stakeholders
- **Respect for human autonomy**: affected stakeholders' capacity to live and act according to their own reasons and motives should be respected

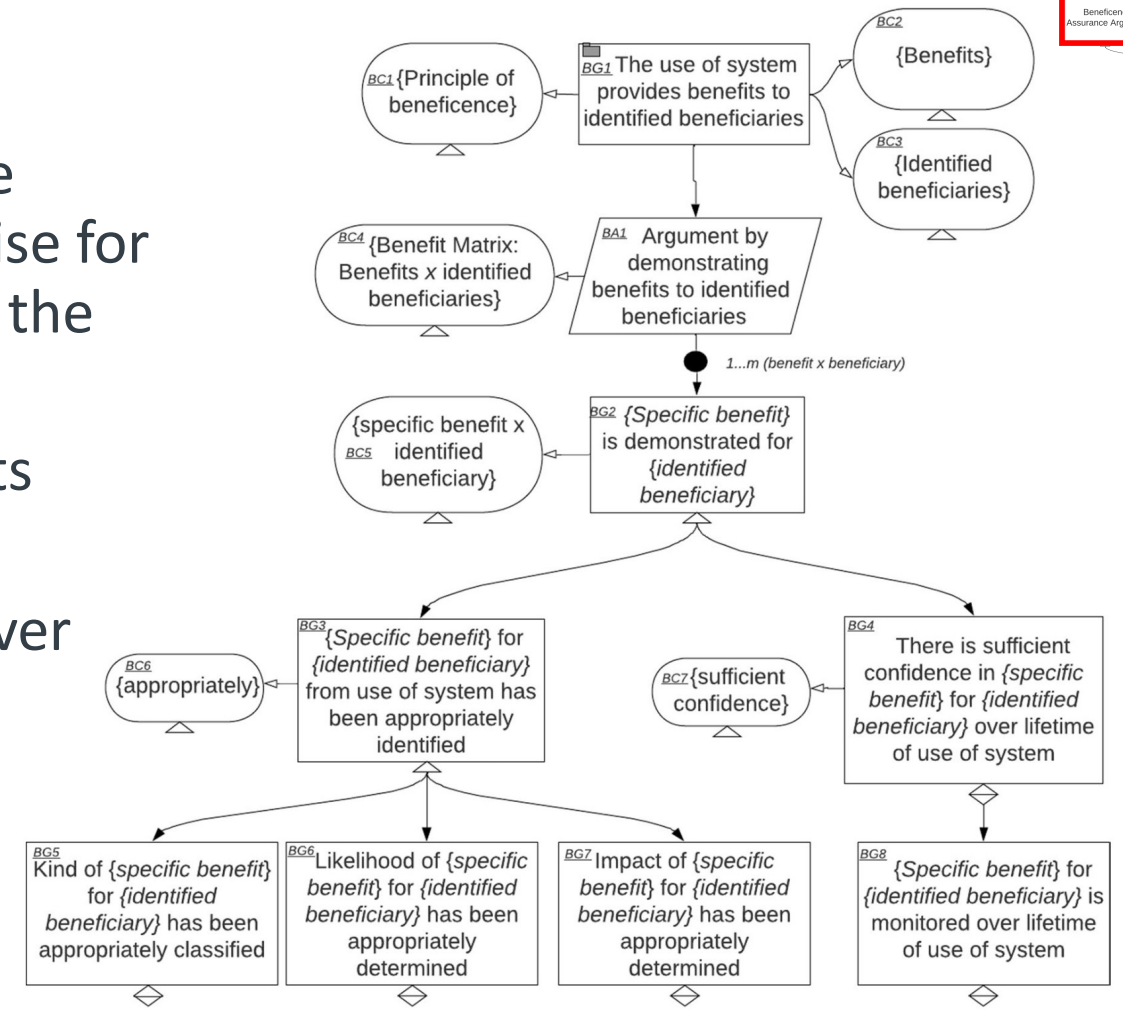
The Ethical Assurance Argument

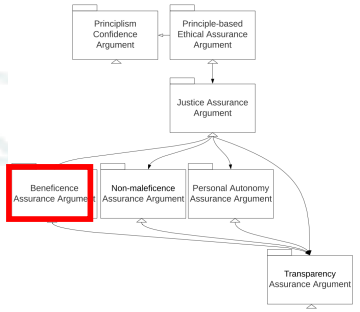


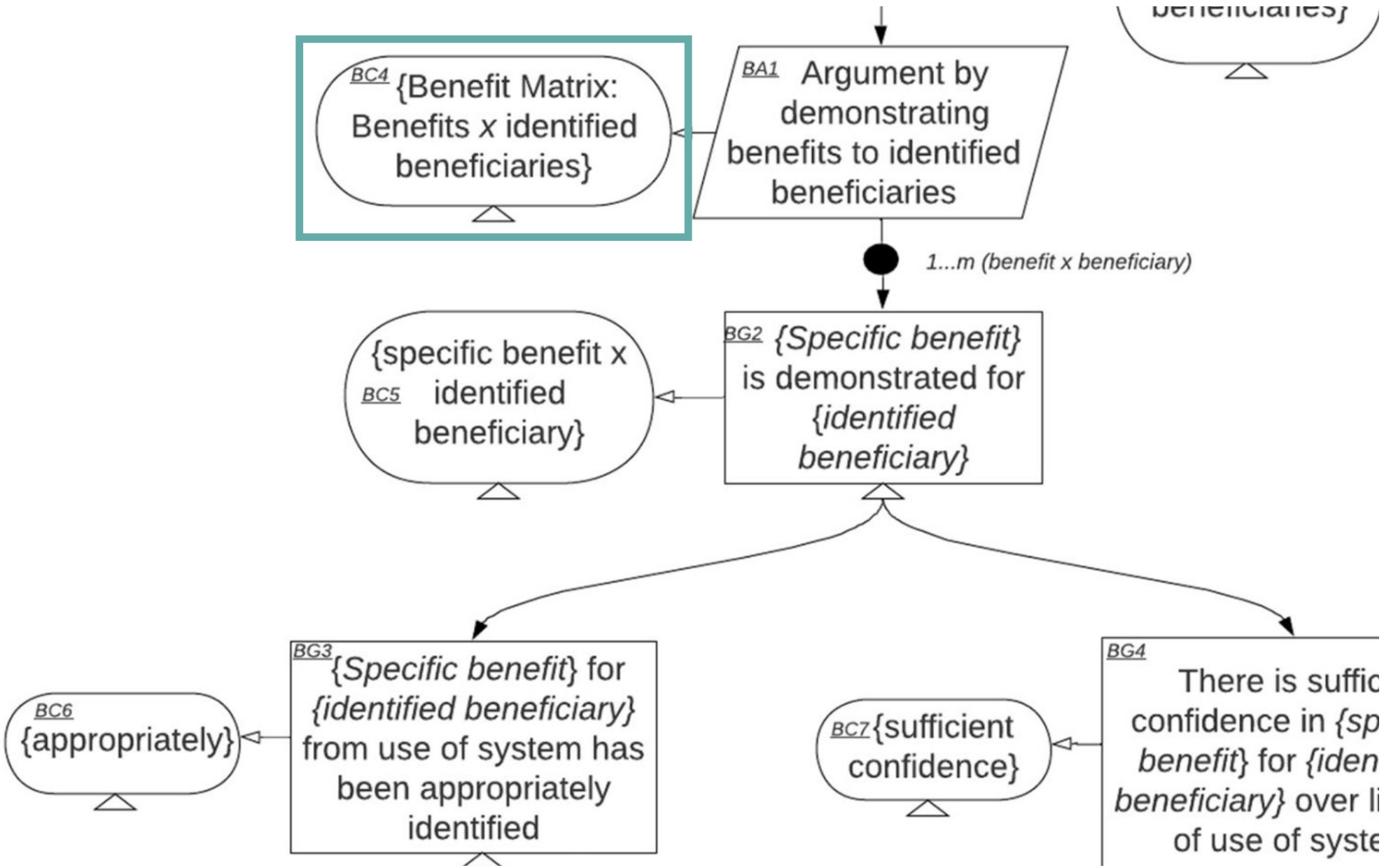
Beneficence argument

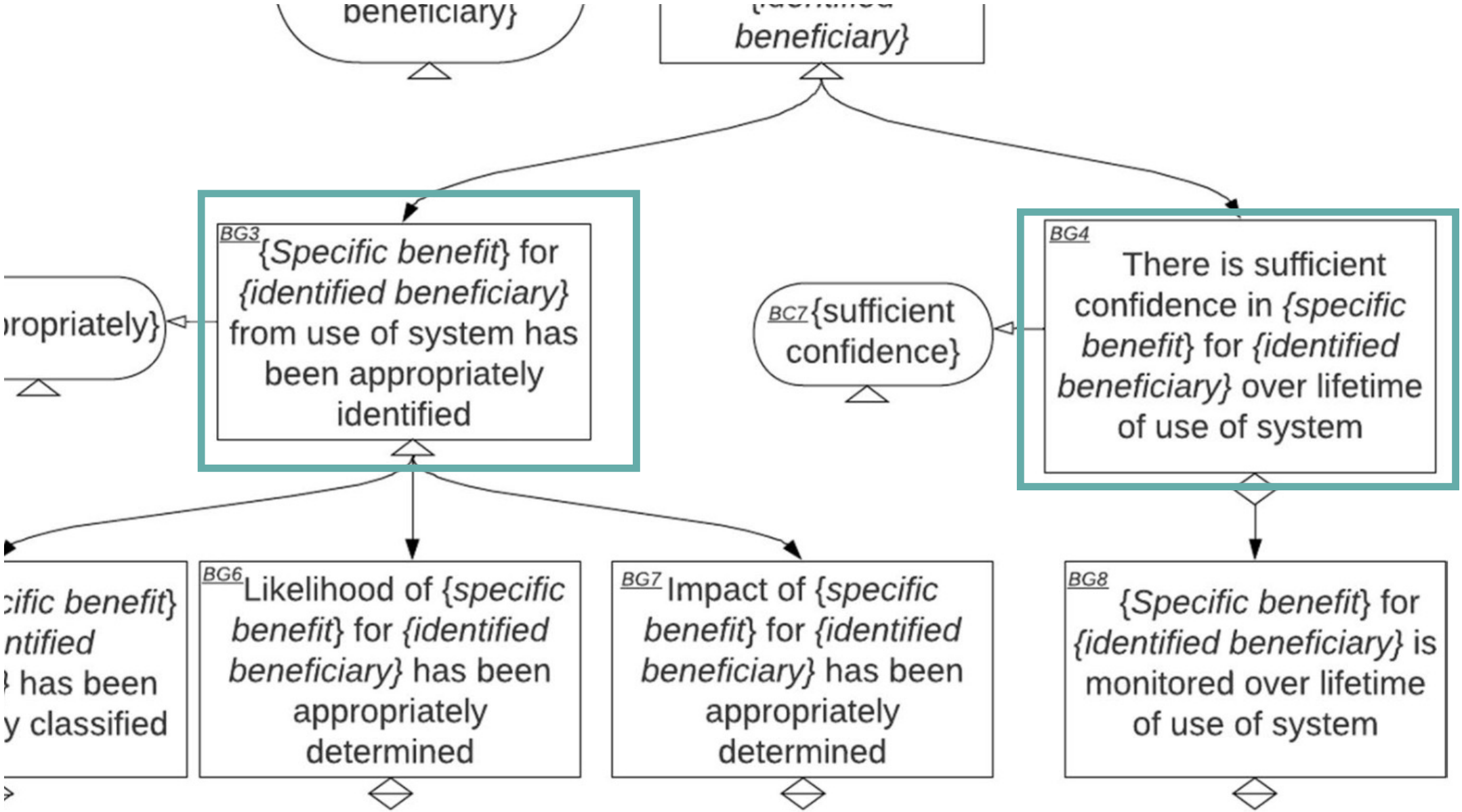
Do good

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





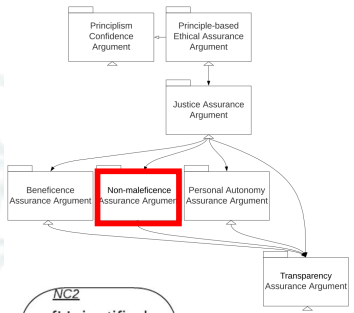

BC4 {Benefit Matrix: Benefits x identified beneficiaries}
BA1 Argument by demonstrating benefits to identified beneficiaries
1...m (benefit x beneficiary)
BC5 {specific benefit x identified beneficiary}
BG2 {Specific benefit} is demonstrated for {identified beneficiary}
BG3 {Specific benefit} for {identified beneficiary} from use of system has been appropriately identified
BC6 {appropriately}
BC7 {sufficient confidence}
BG4
There is suffic
confidence in {sp
benefit} for {iden
beneficiary} over li
of use of syste

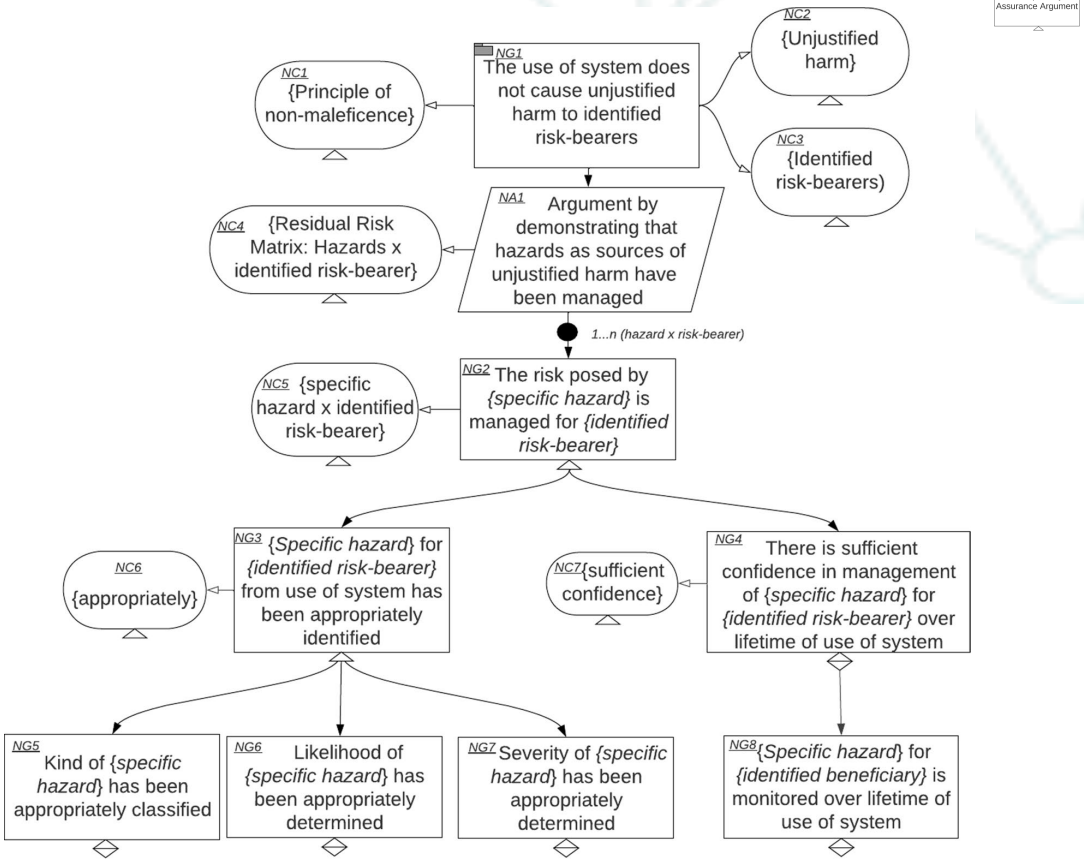

beneficiary}
beneficiary}
BG3 {Specific benefit} for {identified beneficiary} from use of system has been appropriately identified
ropriately}
BC7 {sufficient confidence}
BG4 There is sufficient confidence in {specific benefit} for {identified beneficiary} over lifetime of use of system
cific benefit} ntified has been y classified
BG6 Likelihood of {specific benefit} for {identified beneficiary} has been appropriately determined
BG7 Impact of {specific benefit} for {identified beneficiary} has been appropriately determined
BG8 {Specific benefit} for {identified beneficiary} is monitored over lifetime of use of system

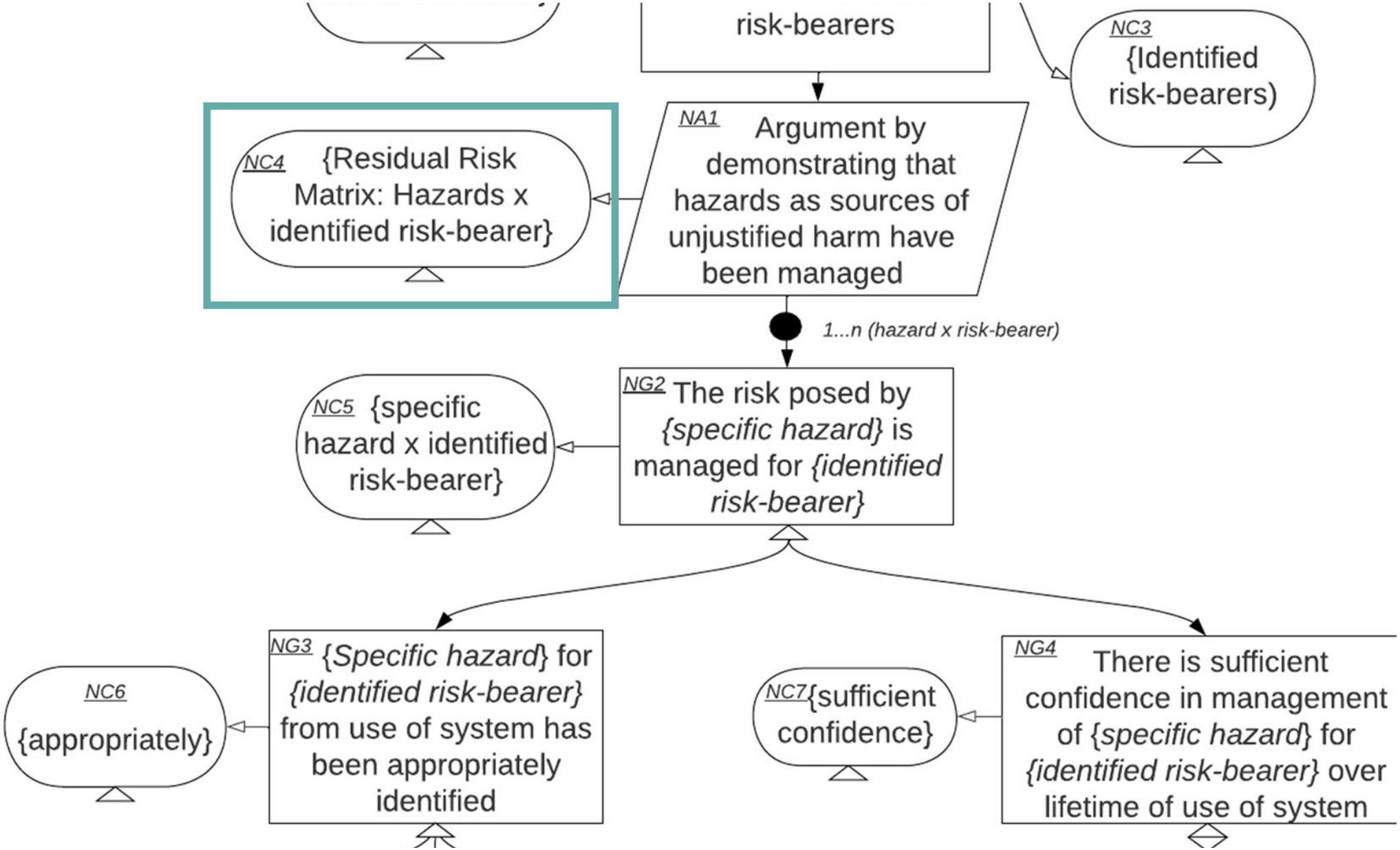
Non-maleficence argument

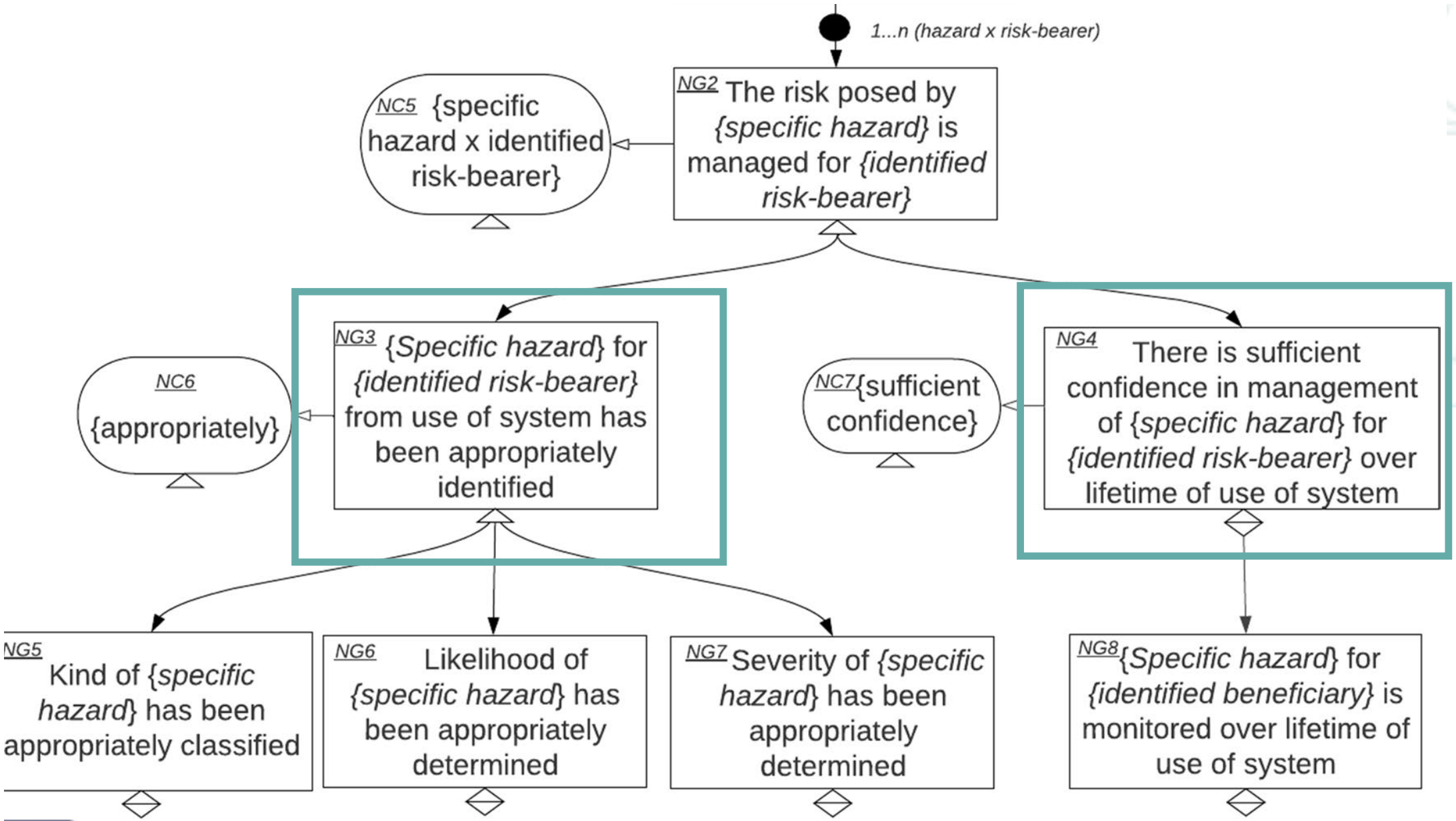
Do no (unjustified) harm

- 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





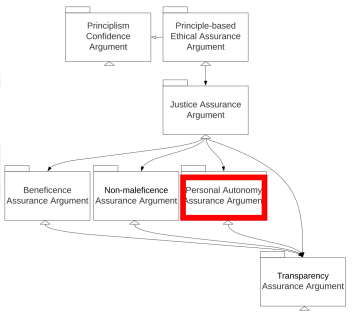

risk-bearers
NC3
{Identified risk-bearers)
NA1
Argument by demonstrating that hazards as sources of unjustified harm have been managed
NC4
{Residual Risk Matrix: Hazards x identified risk-bearer}
1...n (hazard x risk-bearer)
NG2
The risk posed by {specific hazard} is managed for {identified risk-bearer}
NC5
{specific hazard x identified risk-bearer}
NG3
{Specific hazard} for {identified risk-bearer} from use of system has been appropriately identified
NC6
{appropriately}
NG4
There is sufficient confidence in management of {specific hazard} for {identified risk-bearer} over lifetime of use of system
NC7
{sufficient confidence}

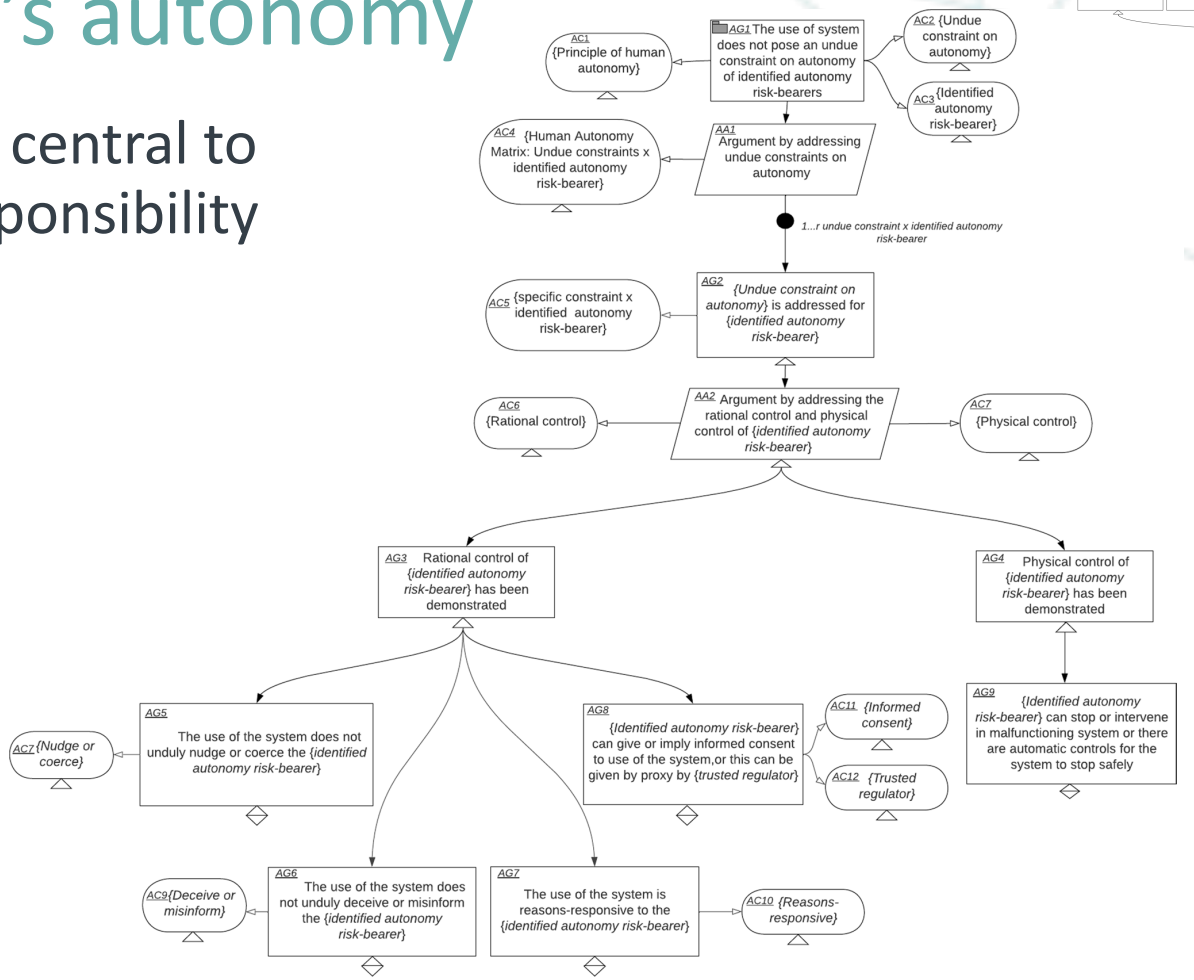

1...n (hazard x risk-bearer)
NG2 The risk posed by {specific hazard} is managed for {identified risk-bearer}
NC5 {specific hazard x identified risk-bearer}
NG3 {Specific hazard} for {identified risk-bearer} from use of system has been appropriately identified
NC6 {appropriately}
NG4 There is sufficient confidence in management of {specific hazard} for {identified risk-bearer} over lifetime of use of system
NC7 {sufficient confidence}
NG5 Kind of {specific hazard} has been appropriately classified
NG6 Likelihood of {specific hazard} has been appropriately determined
NG7 Severity of {specific hazard} has been appropriately determined
NG8 {Specific hazard} for {identified beneficiary} is monitored over lifetime of use of system

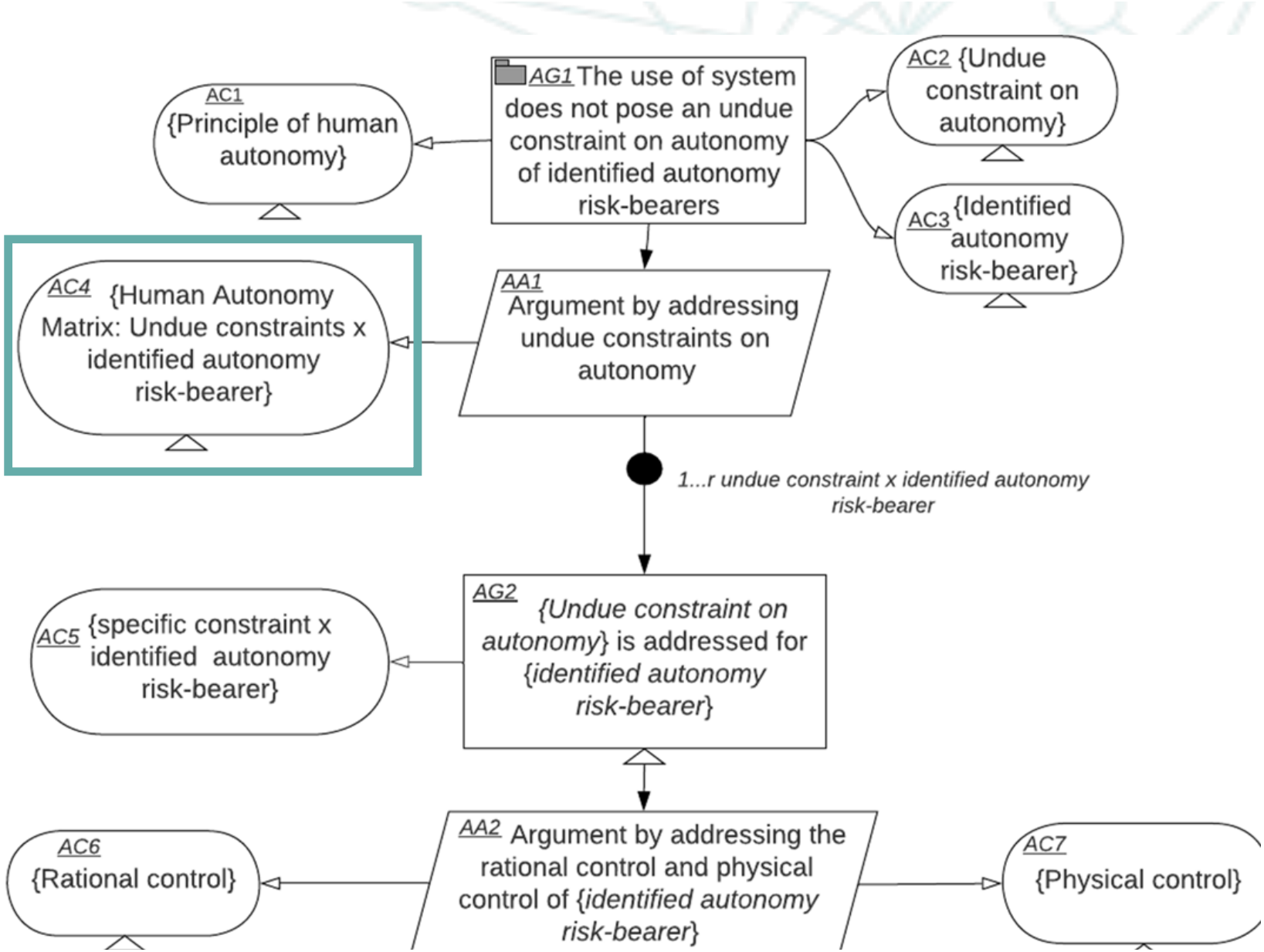
Personal autonomy argument

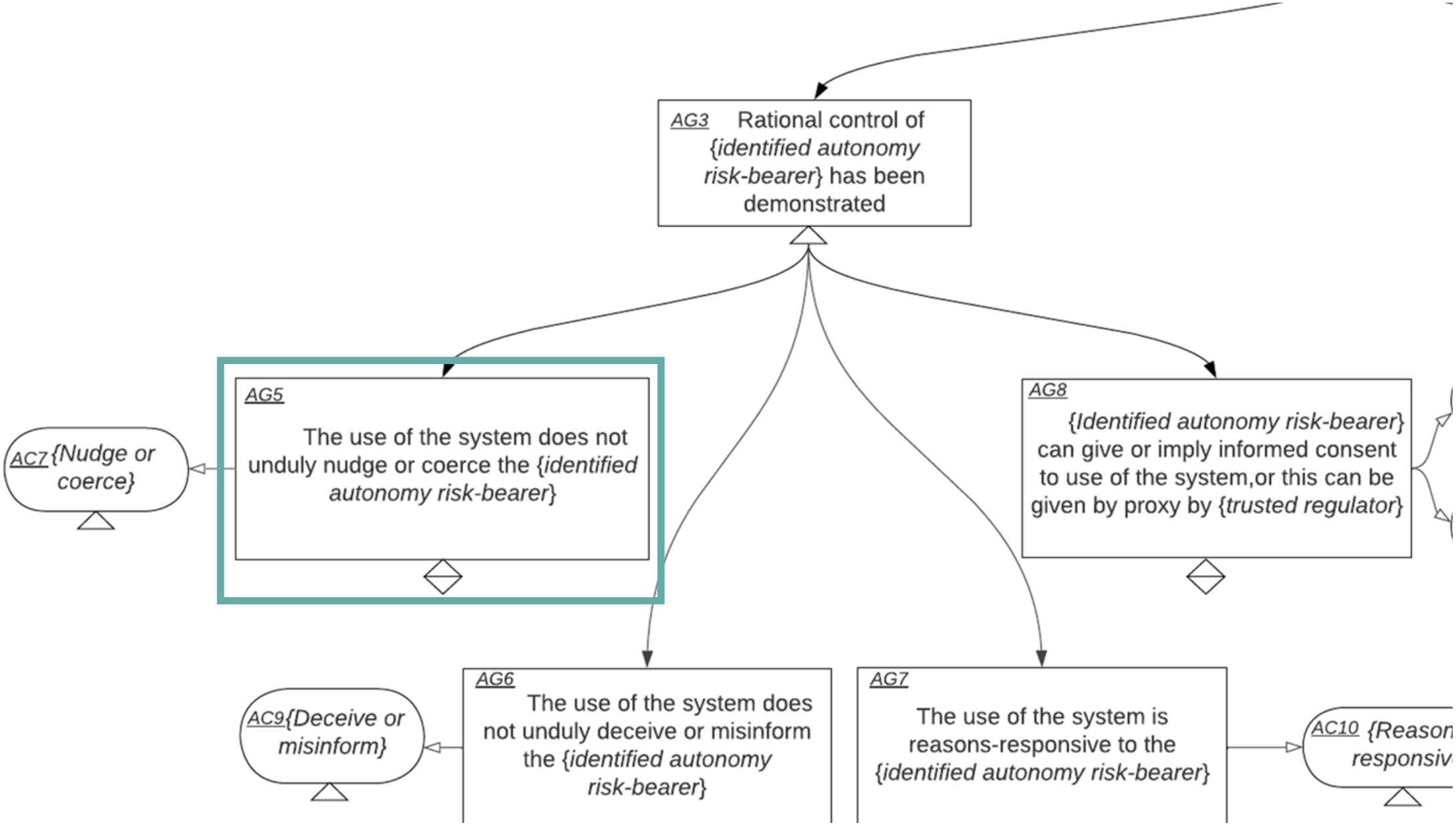
Respect people's autonomy

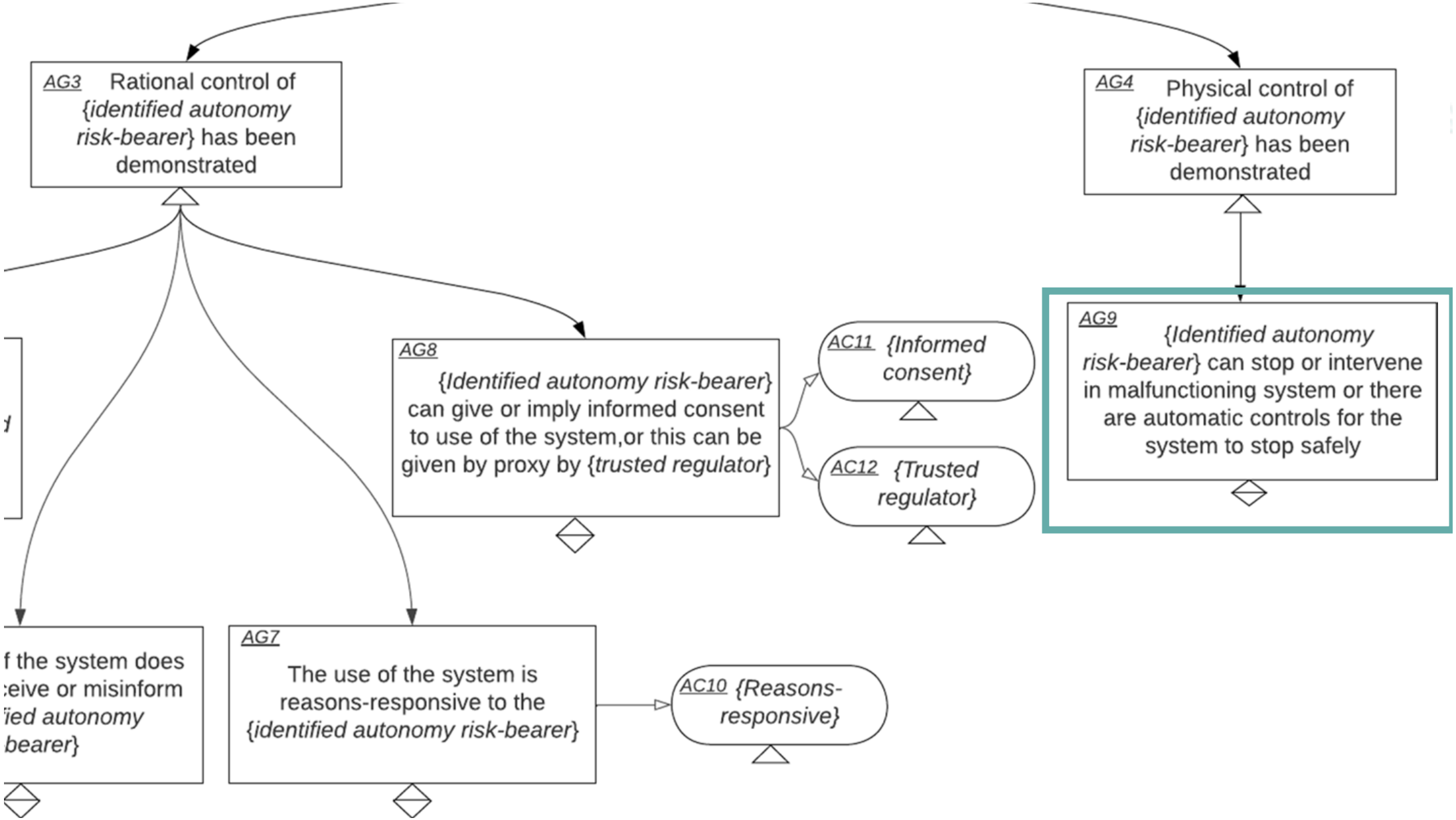
- Personal autonomy is central to moral agency and responsibility



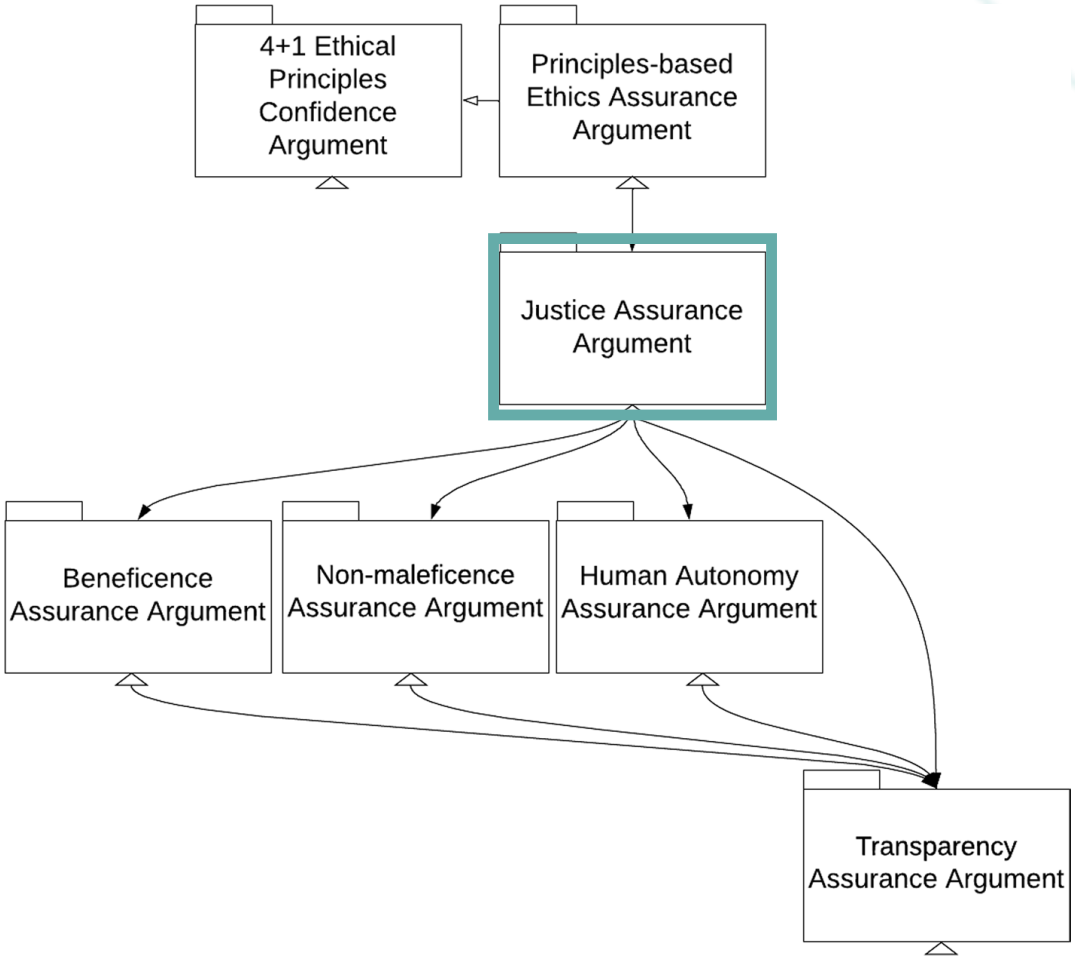



AC1
{Principle of human autonomy}
AG1 The use of system does not pose an undue constraint on autonomy of identified autonomy risk-bearers
AC2 {Undue constraint on autonomy}
AC3 {Identified autonomy risk-bearer}
AC4 {Human Autonomy Matrix: Undue constraints x identified autonomy risk-bearer}
AA1 Argument by addressing undue constraints on autonomy
1...r undue constraint x identified autonomy risk-bearer
AC5 {specific constraint x identified autonomy risk-bearer}
AG2 {Undue constraint on autonomy} is addressed for {identified autonomy risk-bearer}
AC6 {Rational control}
AA2 Argument by addressing the rational control and physical control of {identified autonomy risk-bearer}
AC7 {Physical control}

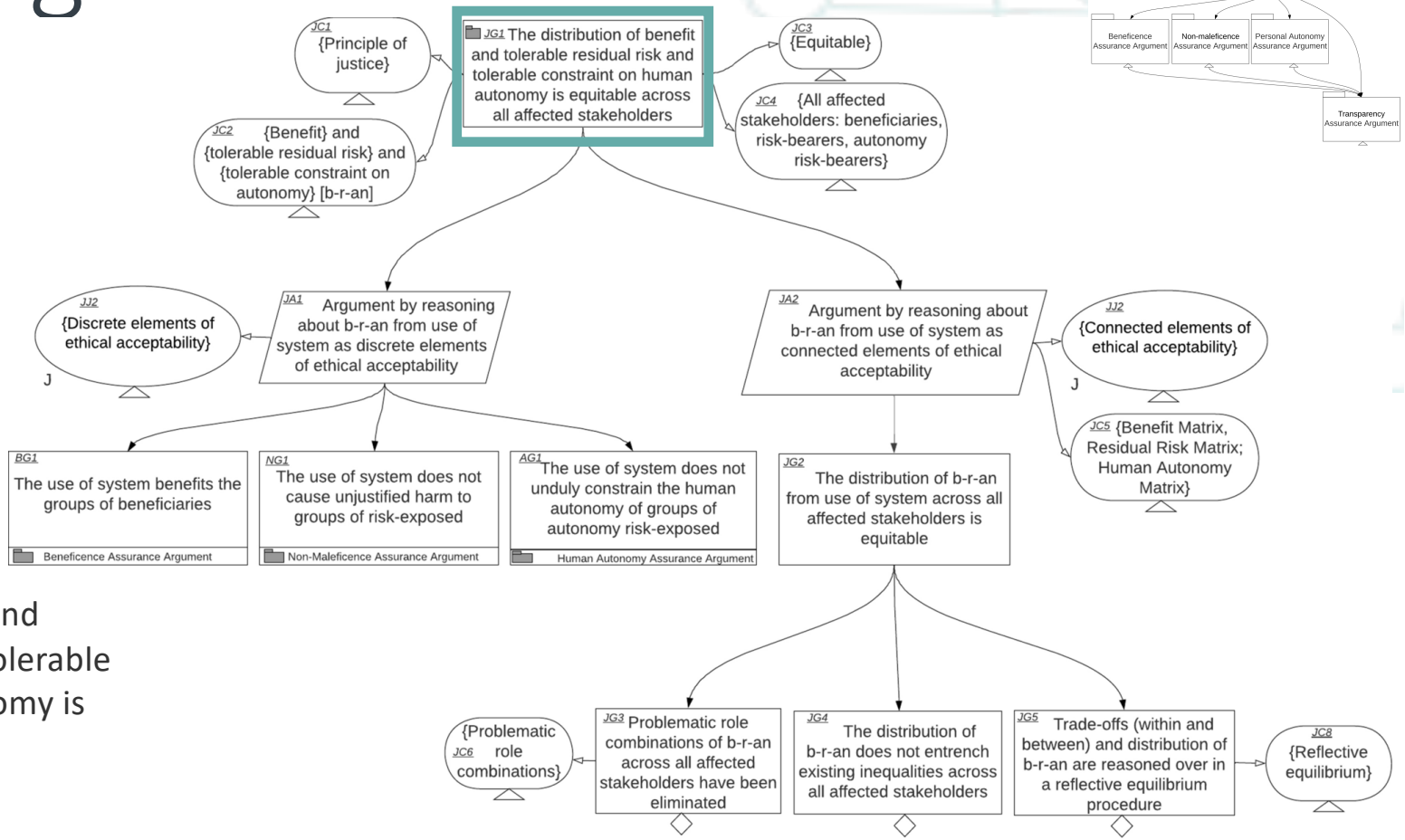

AG3 Rational control of {identified autonomy risk-bearer} has been demonstrated
AG5 The use of the system does not unduly nudge or coerce the {identified autonomy risk-bearer}
AC7 {Nudge or coerce}
AG8 {Identified autonomy risk-bearer} can give or imply informed consent to use of the system,or this can be given by proxy by {trusted regulator}
AG6 The use of the system does not unduly deceive or misinform the {identified autonomy risk-bearer}
AC9 {Deceive or misinform}
AG7 The use of the system is reasons-responsive to the {identified autonomy risk-bearer}
AC10 {Reason
responsiv


AG3 Rational control of {identified autonomy risk-bearer} has been demonstrated
AG4 Physical control of {identified autonomy risk-bearer} has been demonstrated
AG8 {Identified autonomy risk-bearer} can give or imply informed consent to use of the system,or this can be given by proxy by {trusted regulator}
AC11 {Informed consent}
AC12 {Trusted regulator}
AG9 {Identified autonomy risk-bearer} can stop or intervene in malfunctioning system or there are automatic controls for the system to stop safely
f the system does
ceive or misinform
ied autonomy
bearer}
AG7 The use of the system is reasons-responsive to the {identified autonomy risk-bearer}
AC10 {Reasons-responsive}

The Ethical Assurance Argument

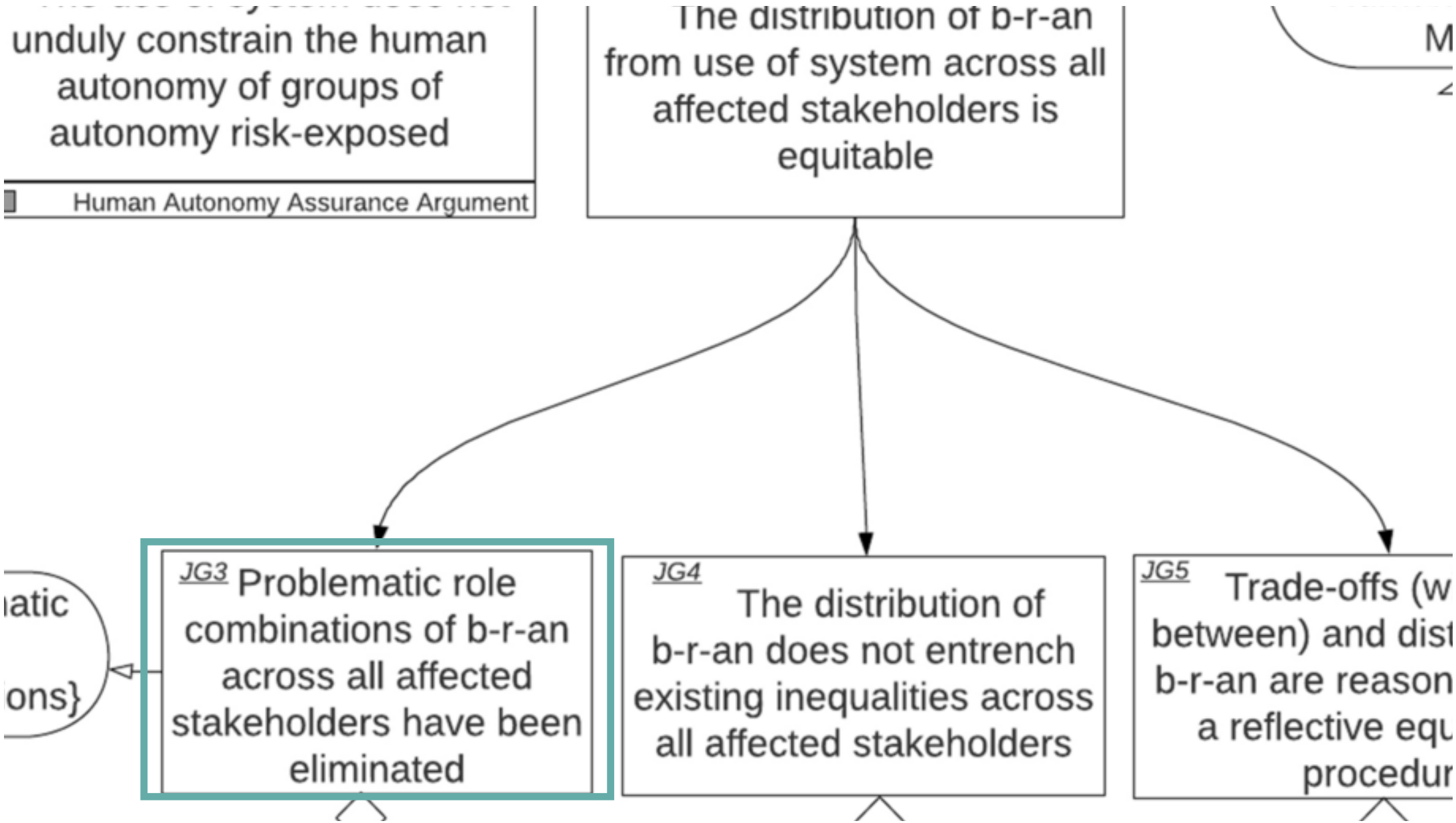


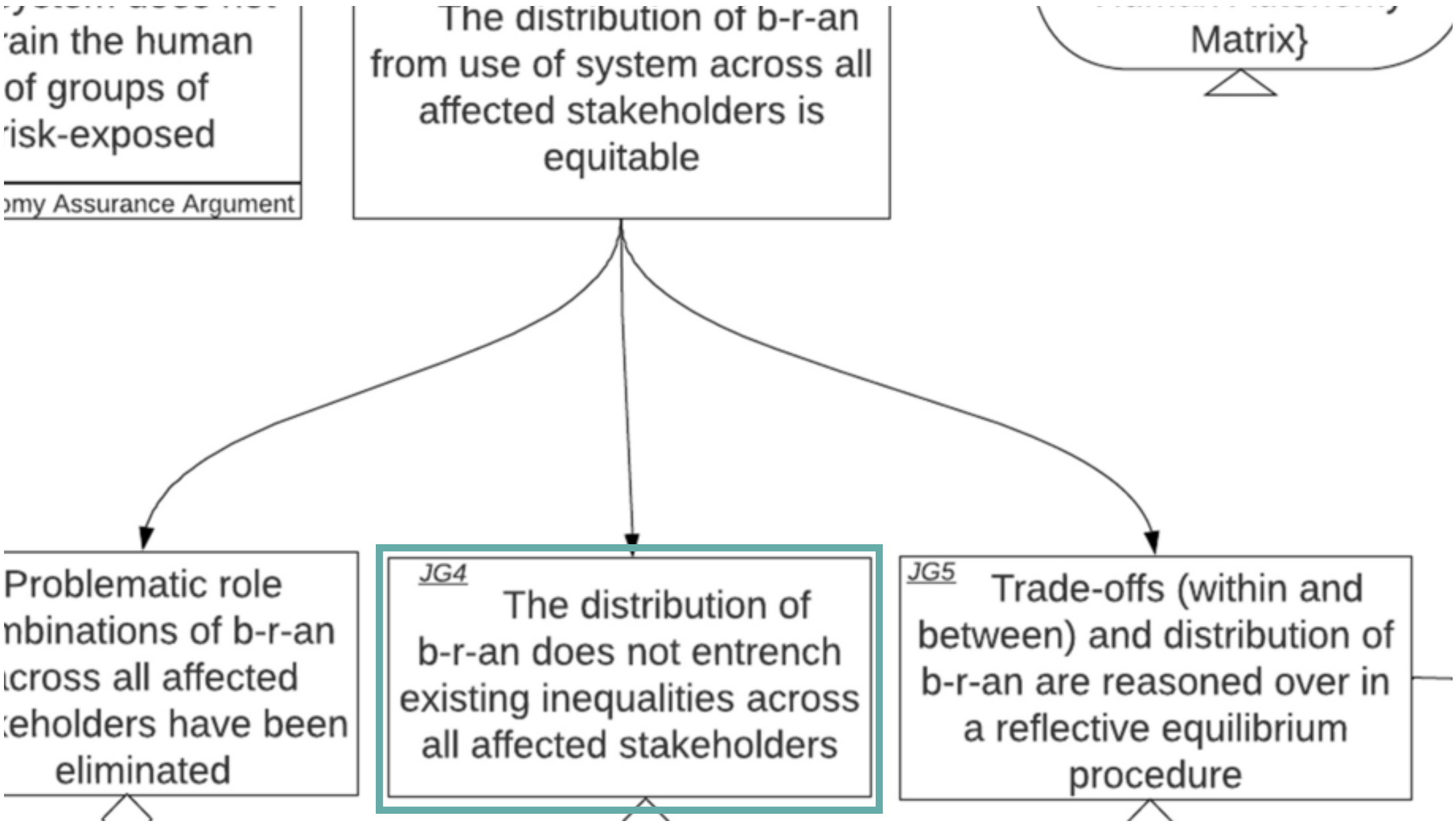
Justice argument

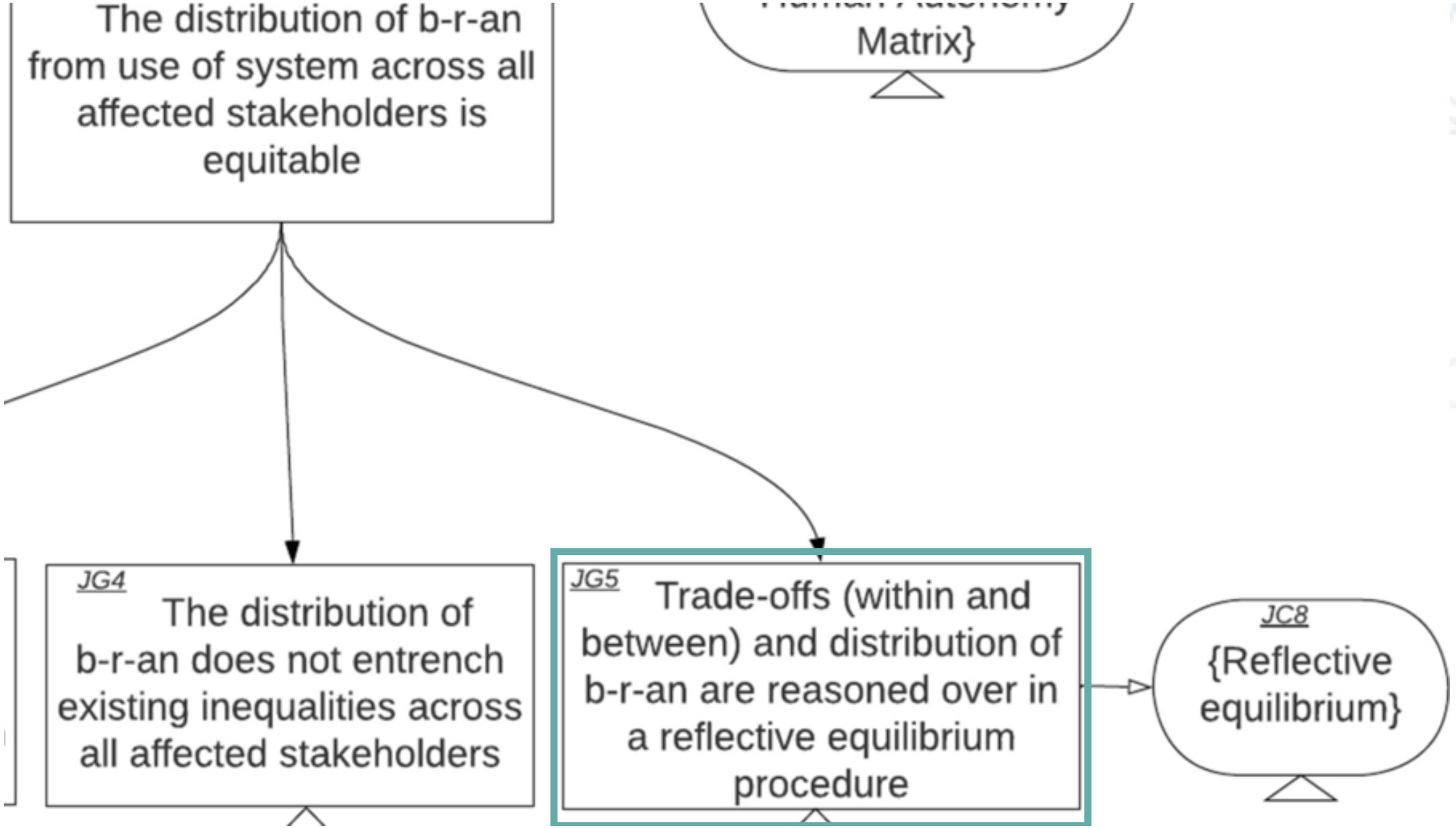


The distribution of benefit and tolerable residual risk and tolerable constraint on human autonomy is equitable across all affected stakeholders

Absence of unacceptable risk of harm caused by the use of AI

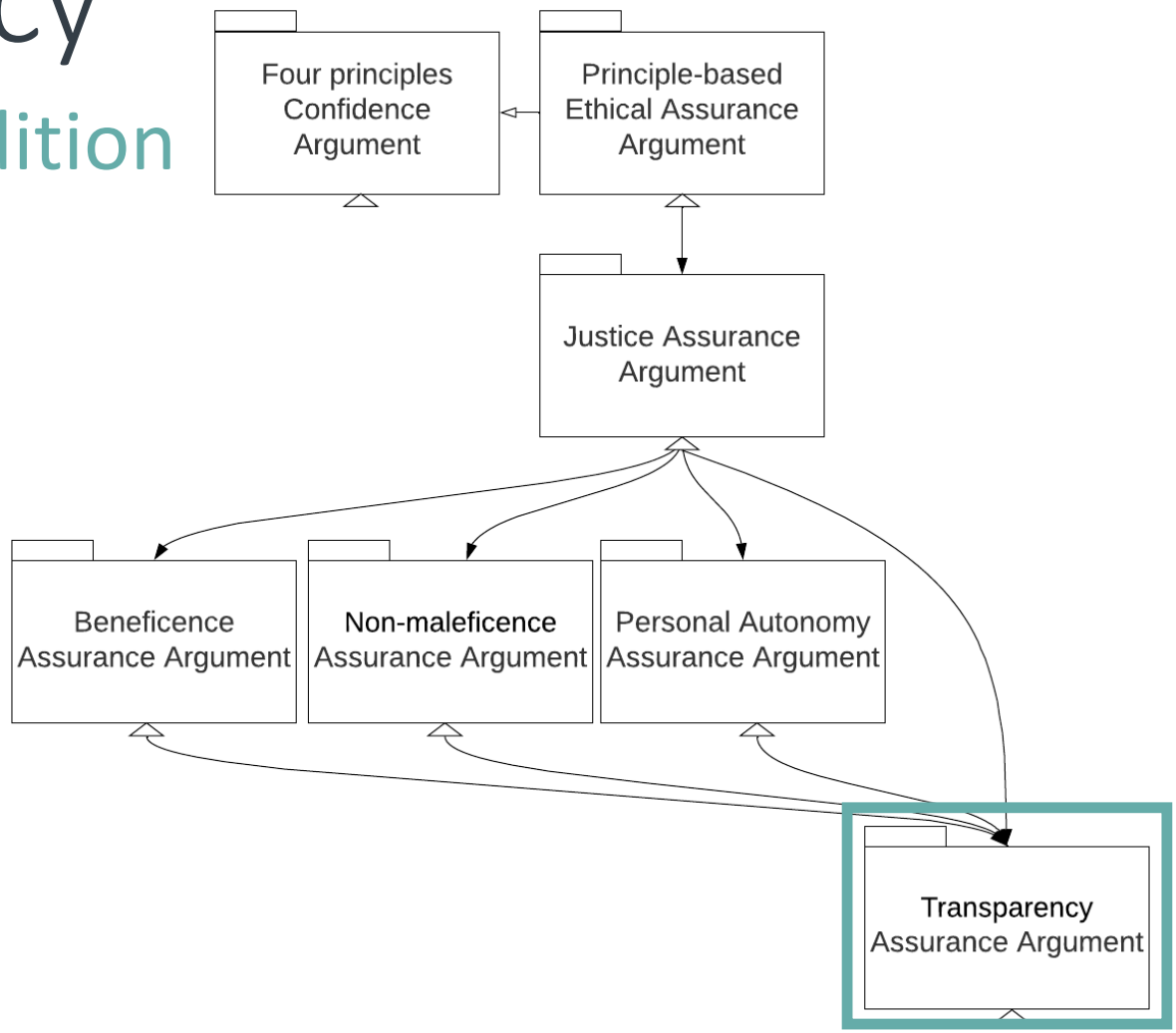

unduly constrain the human autonomy of groups of autonomy risk-exposed
Human Autonomy Assurance Argument
The distribution of b-r-an from use of system across all affected stakeholders is equitable
JG3 Problematic role combinations of b-r-an across all affected stakeholders have been eliminated
JG4 The distribution of b-r-an does not entrench existing inequalities across all affected stakeholders
JG5 Trade-offs (w
between) and dist
b-r-an are reason
a reflective equ
procedur


ain the human
of groups of
isk-exposed
omy Assurance Argument
The distribution of b-r-an from use of system across all affected stakeholders is equitable
Matrix}
Problematic role
nbinations of b-r-an
cross all affected
eholders have been
eliminated
JG4
The distribution of b-r-an does not entrench existing inequalities across all affected stakeholders
JG5
Trade-offs (within and between) and distribution of b-r-an are reasoned over in a reflective equilibrium procedure


Matrix}
The distribution of b-r-an from use of system across all affected stakeholders is equitable
JG4
The distribution of b-r-an does not entrench existing inequalities across all affected stakeholders
JG5
Trade-offs (within and between) and distribution of b-r-an are reasoned over in a reflective equilibrium procedure
JC8
{Reflective equilibrium}

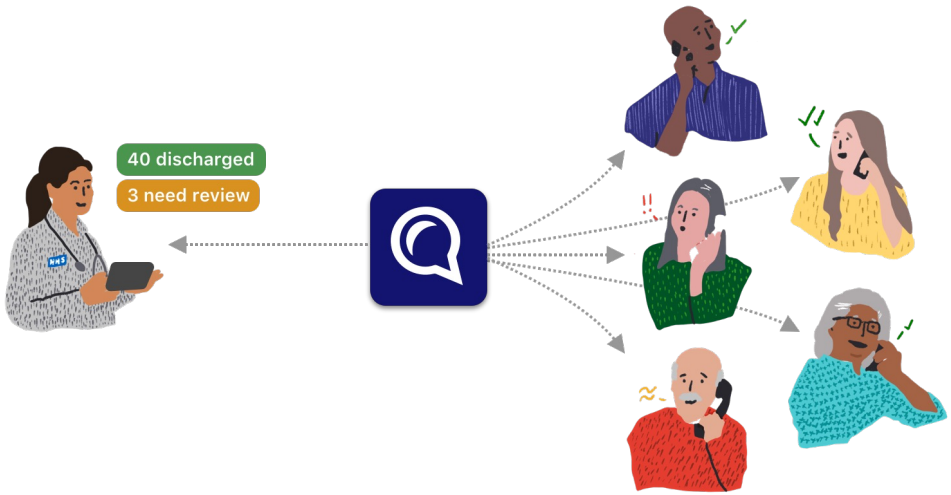
Transparency

An enabling condition



An Example

AI-enabled voice agent for post-operative follow-up



UNIVERSITY of York

Ufonia

Ethics in conversation

Building an ethics assurance case for autonomous AI-enabled voice agents in healthcare

Marten H. L. Kaas
University of York
marten.kaas@york.ac.uk

Zoe Porter
University of York
zoe.porter@york.ac.uk

Ernest Lim
Ufonia Limited
el@ufonia.co

Aisling Higham
Ufonia Limited
ah@ufonia.co

Sarah Khavandi
Ufonia Limited
sk@ufonia.co

Ibrahim Habli
University of York
ibrahim.habli@york.ac.uk

ABSTRACT

The deployment and use of AI systems should be both safe and broadly ethically acceptable. The principles-based ethics assurance argument pattern is one proposal in the AI ethics landscape that seeks to support and achieve that aim. The purpose of this argument pattern or framework is to structure reasoning about, and to communicate and foster confidence in, the ethical acceptability of uses of specific real-world AI systems in complex socio-technical contexts. This paper presents the interim findings of a case study applying this ethics assurance framework to the use of Dora, an AI-based telemedicine system, to assess its viability and usefulness as an approach. The case study process to date has revealed some of the positive ethical impacts of the Dora platform, as well as unexpected insights and areas to prioritise for evaluation, such as risks to the frontline clinician, particularly in respect of clinician autonomy. The ethics assurance argument pattern offers a practical framework not just for identifying issues to be addressed, but also to start to construct solutions in the form of adjustments to the distribution of benefits, risks and constraints on human autonomy that could reduce ethical disparities across affected stakeholders. Though many challenges remain, this research represents a step in the direction towards the development and use of safe and ethically acceptable AI systems and, ideally, a shift towards more comprehensive and inclusive evaluations of AI systems in general.

CCS CONCEPTS

• **general and reference**; • **document types**; • **general conference proceedings**;

KEYWORDS

ethics assurance, case study, AI-based telemedicine, Dora platform, medical device, ethical acceptability

ACM Reference Format:
Marten H. L. Kaas, Zoe Porter, Ernest Lim, Aisling Higham, Sarah Khavandi, and Ibrahim Habli. 2023. Ethics in conversation: Building an ethics assurance case for autonomous AI-enabled voice agents in healthcare. In *First International Symposium on Trustworthy Autonomous Systems (TAS '23), July 11, 12, 2023, Edinburgh, United Kingdom.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3597512.3599713

TAS '23, July 11, 12, 2023, Edinburgh, United Kingdom

ACM ISBN 979-8-4007-0734-6/23/07…$15.00
https://doi.org/10.1145/3597512.3599713

1 INTRODUCTION

As AI-based systems increasingly permeate society, it is widely recognized that new approaches to ensuring the safety and efficacy of such systems are needed. But merely ensuring the safety of AI-based systems is not enough. The human tendency to defer to suggestions generated by AI systems, their "black box" and dynamically updating nature, gaps in regulation and an emphasis on being first to market all conspire to threaten not just the safe deployment and use of AI systems, but their ethical acceptability as well. This paper attempts to address the gap between meeting minimum safety requirements and ethical acceptability by evaluating the plausibility, viability and value of instantiating the ethics assurance argument pattern proposed by Porter et al. [41] in the healthcare context for an AI-based telemedicine system. Our interest is not only in safety, but rather something more ambitious: ethical acceptability. As impressive as AI systems are, their abilities are still derived from humans and as such lack the sort of normative commitments and capacity for considered judgement that humans have [47]. It therefore falls on us, the developers, investors, regulators, users, researchers and affected stakeholders, to carefully consider the consequences of deploying AI systems. Our research is, we maintain, one step towards ensuring the responsible development of AI systems whose impacts can be difficult to predict, far-reaching and long lasting.

This paper is structured as follows. In section 2, we introduce the system, Dora, and describe its place in the clinical pathway as well as the regulatory landscape governing its use. In section 3, we describe the principles-based ethics assurance argument pattern and in section 4 apply the argument pattern to Dora and explain our preliminary results. Lastly, in section 5 we draw out some conclusions of our research including limitations of our work and areas for future research.

2 THE TECHNOLOGY (DORA) AND ITS CONTEXT

2.1 Introduction to Dora

Healthcare is facing a workforce crisis. In the UK, demand on the National Health Service (NHS) is increasing beyond the current capacity of healthcare staff [50]. With increasing demands, and a

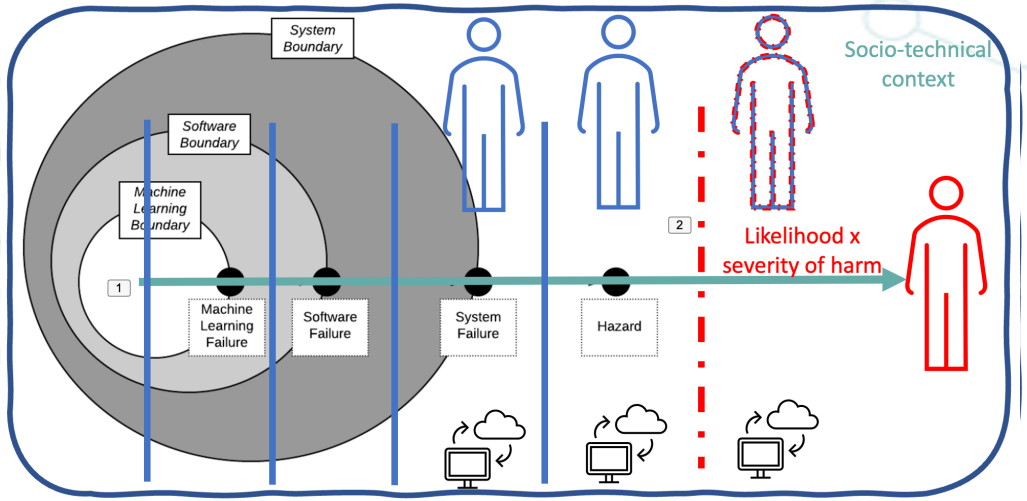
Respect for Human Autonomy

Clinicians' perspectives

Impact professional competence

Allocation of legal liability

Impact on psychological well-being



Ethics in conversation

Building an ethics assurance case for autonomous AI-enabled voice agents in healthcare

Marten H. L. Kaas
University of York
marten.kaas@york.ac.uk

Zoe Porter
University of York
zoe.porter@york.ac.uk

Ernest Lim
Ufonia Limited
el@ufonia.co

Aisling Higham
Ufonia Limited
ah@ufonia.co

Sarah Khavandi
Ufonia Limited
sk@ufonia.co

Ibrahim Habli
University of York
ibrahim.habli@york.ac.uk

ABSTRACT

The deployment and use of AI systems should be both safe and broadly ethically acceptable. The principles-based ethics assurance argument pattern is one proposal in the AI ethics landscape that seeks to support and achieve that aim. The purpose of this argument pattern or framework is to structure reasoning about, and to communicate and foster confidence in, the ethical acceptability of uses of specific real-world AI systems in complex socio-technical contexts. This paper presents the interim findings of a case study applying this ethics assurance framework to the use of Dora, an AI-based telemedicine system, to assess its viability and usefulness as an approach. The case study process to date has revealed some of the positive ethical impacts of the Dora platform, as well as unexpected insights and areas to prioritise for evaluation, such as risks to the frontline clinician, particularly in respect of clinician autonomy. The ethics assurance argument pattern offers a practical framework not just for identifying issues to be addressed, but also to start to construct solutions in the form of adjustments to the distribution of benefits, risks and constraints on human autonomy that could reduce ethical disparities across affected stakeholders. Though many challenges remain, this research represents a step in the direction towards the development and use of safe and ethically acceptable AI systems and, ideally, a shift towards more comprehensive and inclusive evaluations of AI systems in general.

CCS CONCEPTS

• **general and reference**; • **document types**; • **general conference proceedings**;

KEYWORDS

ethics assurance, case study, AI-based telemedicine, Dora platform, medical device, ethical acceptability

ACM Reference Format:

Marten H. L. Kaas, Zoe Porter, Ernest Lim, Aisling Higham, Sarah Khavandi, and Ibrahim Habli. 2023. Ethics in conversation: Building an ethics assurance case for autonomous AI-enabled voice agents in healthcare. In *First International Symposium on Trustworthy Autonomous Systems (TAS '23), July 11, 12, 2023, Edinburgh, United Kingdom.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3597512.3599713

TAS '23, July 11, 12, 2023, Edinburgh, United Kingdom

ACM ISBN 979-8-4007-0734-6/23/07…$15.00
https://doi.org/10.1145/3597512.3599713

1 INTRODUCTION

As AI-based systems increasingly permeate society, it is widely recognized that new approaches to ensuring the safety and efficacy of such systems are needed. But merely ensuring the safety of AI-based systems is not enough. The human tendency to defer to suggestions generated by AI systems, their "black box" and dynamically updating nature, gaps in regulation and an emphasis on being first to market all conspire to threaten not just the safe deployment and use of AI systems, but their ethical acceptability as well. This paper attempts to address the gap between meeting minimum safety requirements and ethical acceptability by evaluating the plausibility, viability and value of instantiating the ethics assurance argument pattern proposed by Porter et al. [41] in the healthcare context for an AI-based telemedicine system. Our interest is not only in safety, but rather something more ambitious: ethical acceptability. As impressive as AI systems are, their abilities are still derived from humans and as such lack the sort of normative commitments and capacity for considered judgement that humans have [47]. It therefore falls on us, the developers, investors, regulators, users, researchers and affected stakeholders, to carefully consider the consequences of deploying AI systems. Our research is, we maintain, one step towards ensuring the responsible development of AI systems whose impacts can be difficult to predict, far-reaching and long lasting.

This paper is structured as follows. In section 2, we introduce the system, Dora, and describe its place in the clinical pathway as well as the regulatory landscape governing its use. In section 3, we describe the principles-based ethics assurance argument pattern and in section 4 apply the argument pattern to Dora and explain our preliminary results. Lastly, in section 5 we draw out some conclusions of our research including limitations of our work and areas for future research.

2 THE TECHNOLOGY (DORA) AND ITS CONTEXT

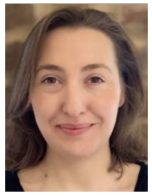
2.1 Introduction to Dora

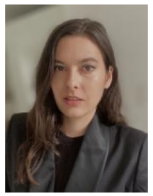
Healthcare is facing a workforce crisis. In the UK, demand on the National Health Service (NHS) is increasing beyond the current capacity of healthcare staff [50]. With increasing demands, and a

What’s next?

Establishing responsibility

AR-TAS: Assuring Responsibility for Trustworthy Autonomous Systems



Lloyd's Register Foundation

Ufonia

NHS Digital

MIRA

SHEFFIELD ROBOTICS

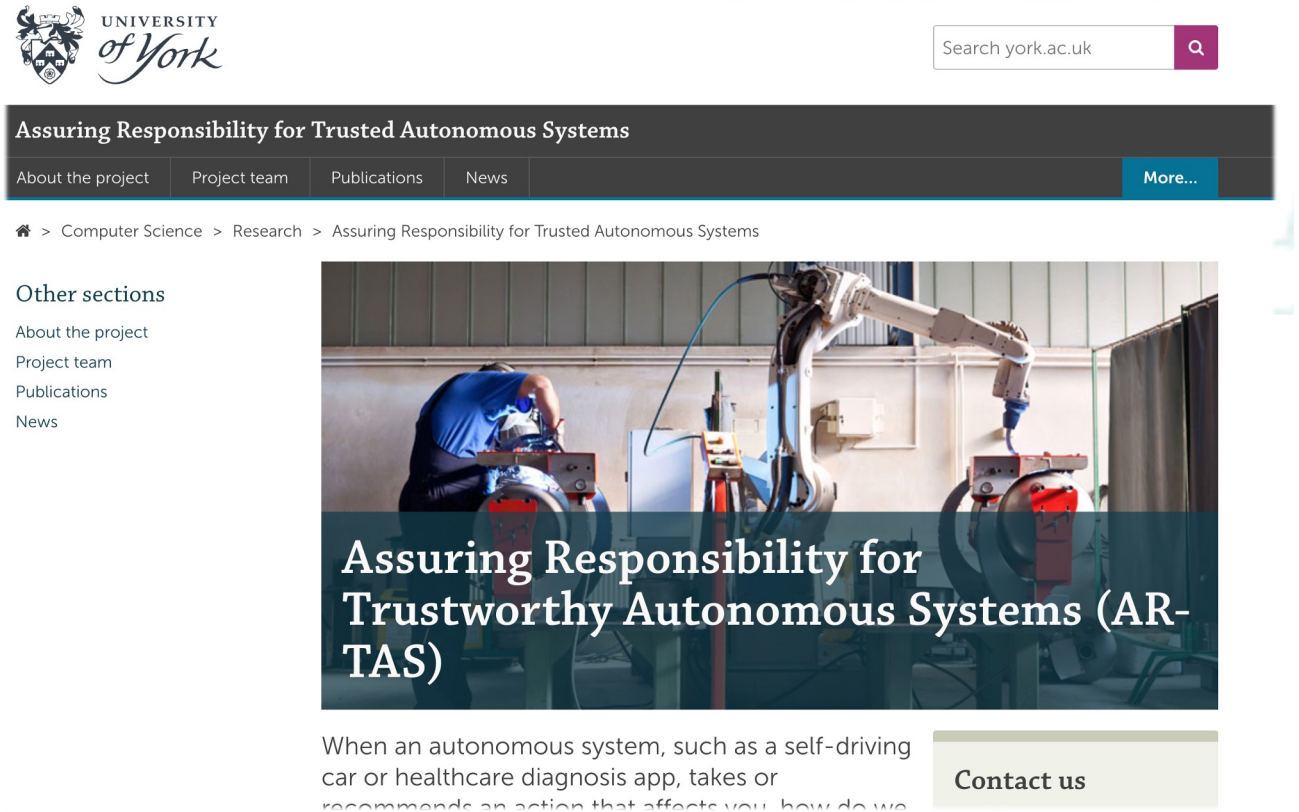
WAYVE

JAGUAR

LAND ROVER

NASA

Project Reference:	Funded Period:	Funded Value:
EP/W011239/1	Jan 22 - Jun 24	£703,615

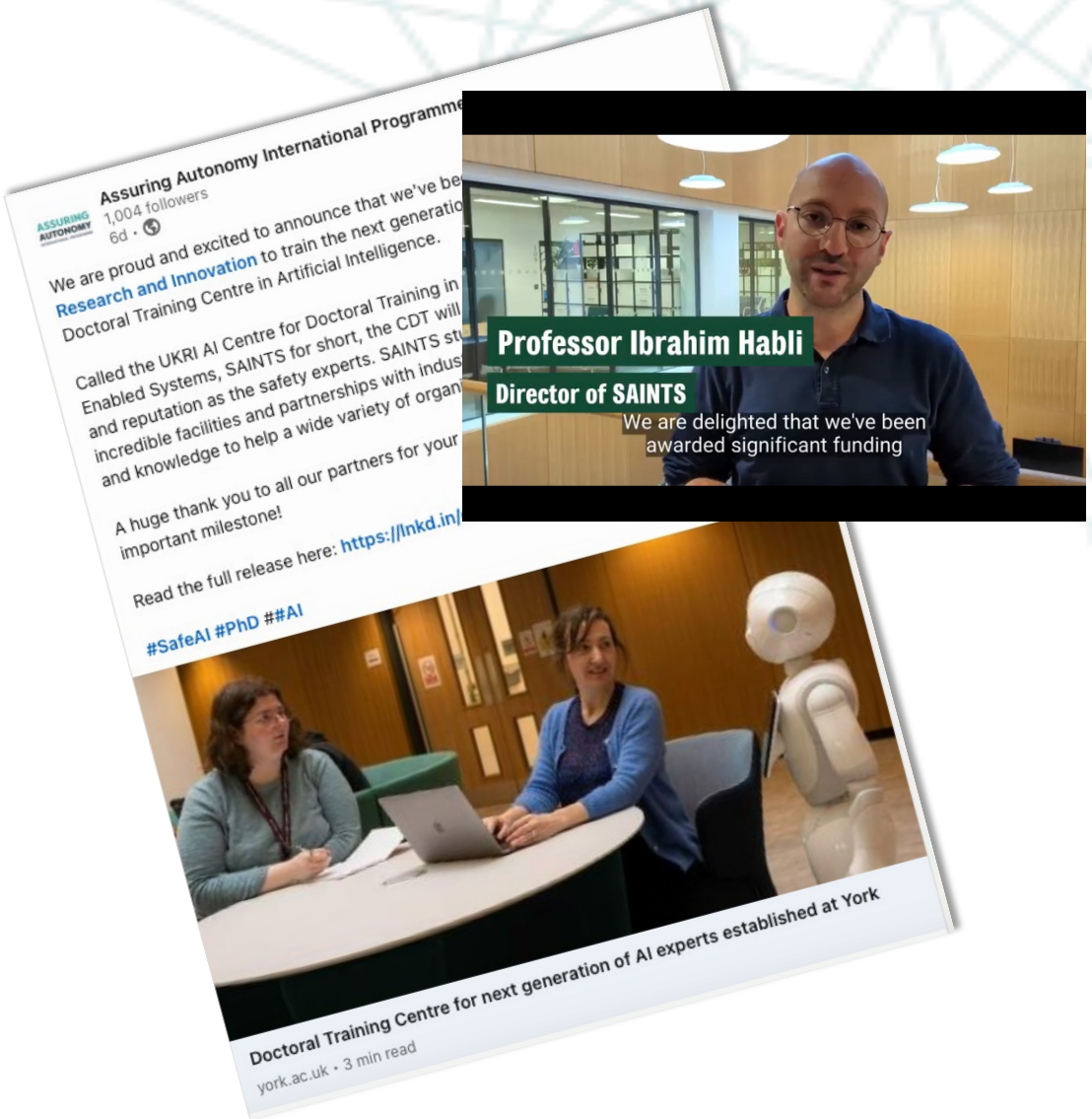


https://www.cs.york.ac.uk/research/trusted-autonomous-systems/

What else?

SAINTS AI Safety Centre

- *The UK's only Centre for Doctoral Training in Safe AI*
  - 60 PhD students
  - 34 industry/regulatory partners
  - £16.2M investment
  - Focus on the *Lifelong Safety Assurance of AI-Enabled Autonomous Systems*
  - First cohort: October 2024
  - Industry, policy, regulatory & academic careers



https://www.york.ac.uk/news-and-events/news/2023/quality/doctoral-training-centre-ai-safety/

UNIVERSITY of York

ASSURING AUTONOMY

INTERNATIONAL PROGRAMME