
The Alan Turing Institute
Day 5
Responsible Communication
10101001

Recap

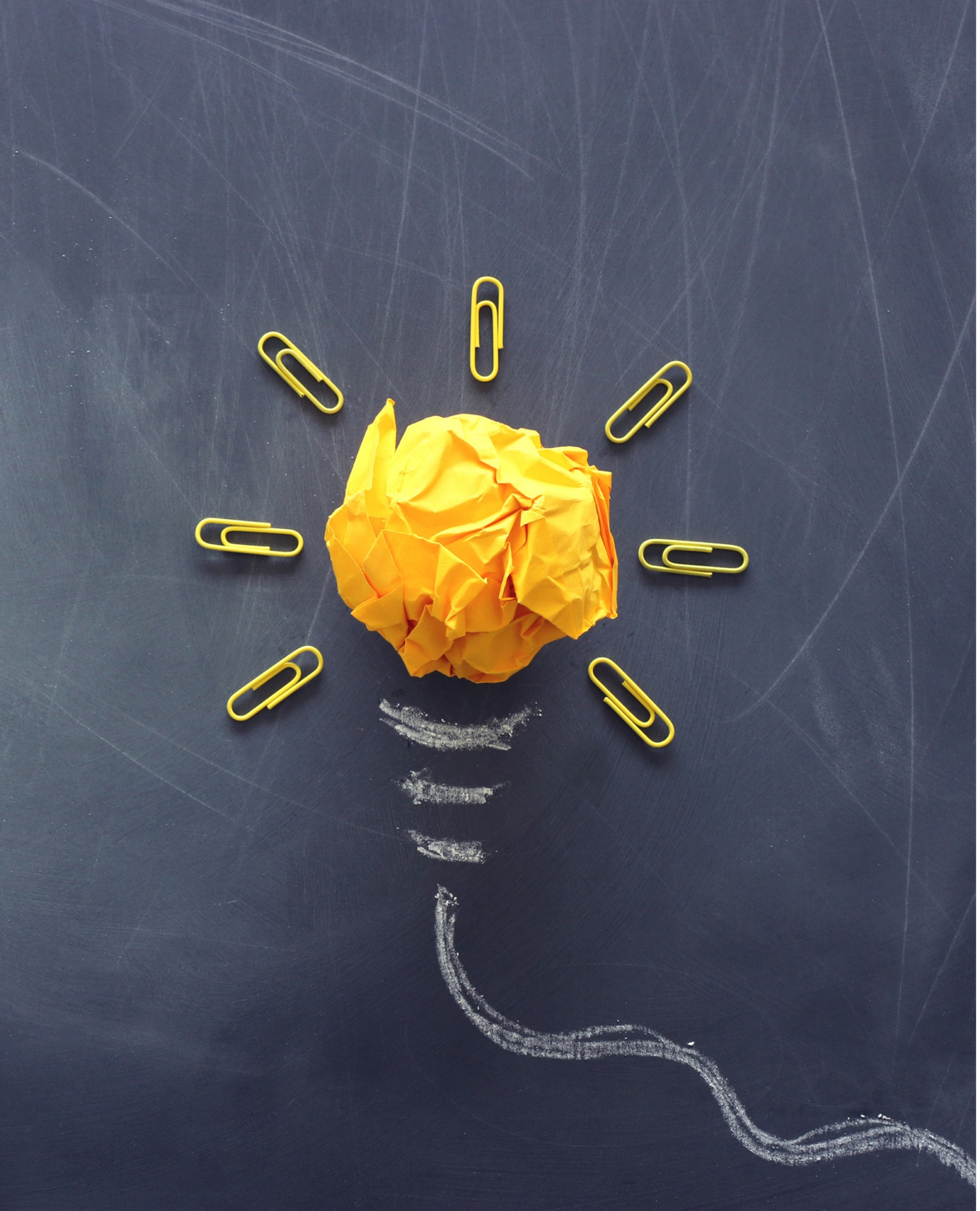
Day 4

- Model Development
  - Preprocessing & Feature Engineering
  - Model Selection & Training
  - Model Testing & Validation
  - Model Reporting
- System Deployment
  - Model Productionalisation
  - User Training
  - System Use & monitoring
  - Model Updating or Deprovisioning
- Guest Lecture - Professor Sabina Leonelli

Overview

Day 5

- What is Argument-Based Assurance?
- Assurance and Responsible Communication
- Goals, Properties, and Evidence
- Wrap-Up



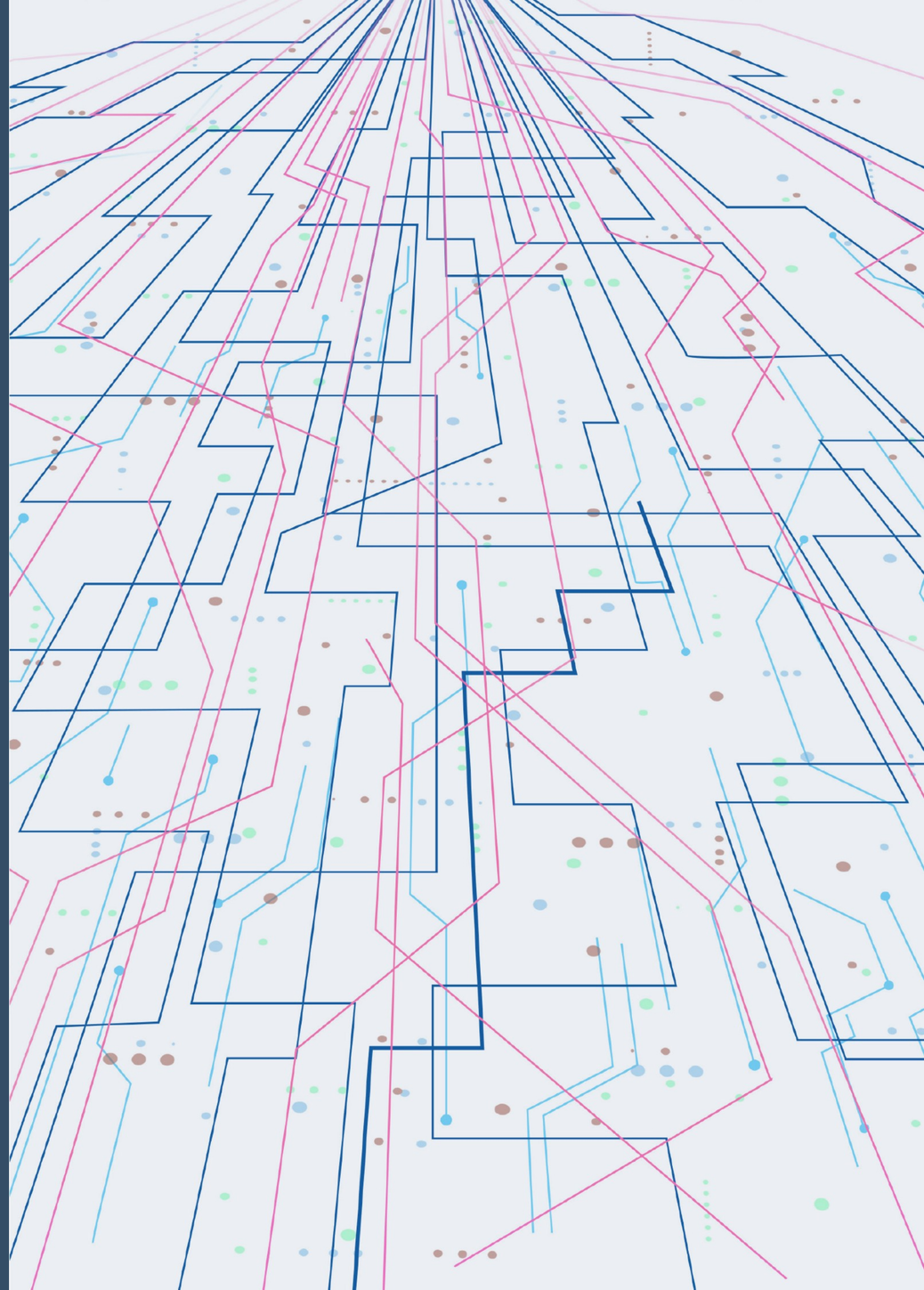
Day 5

Learning Objectives

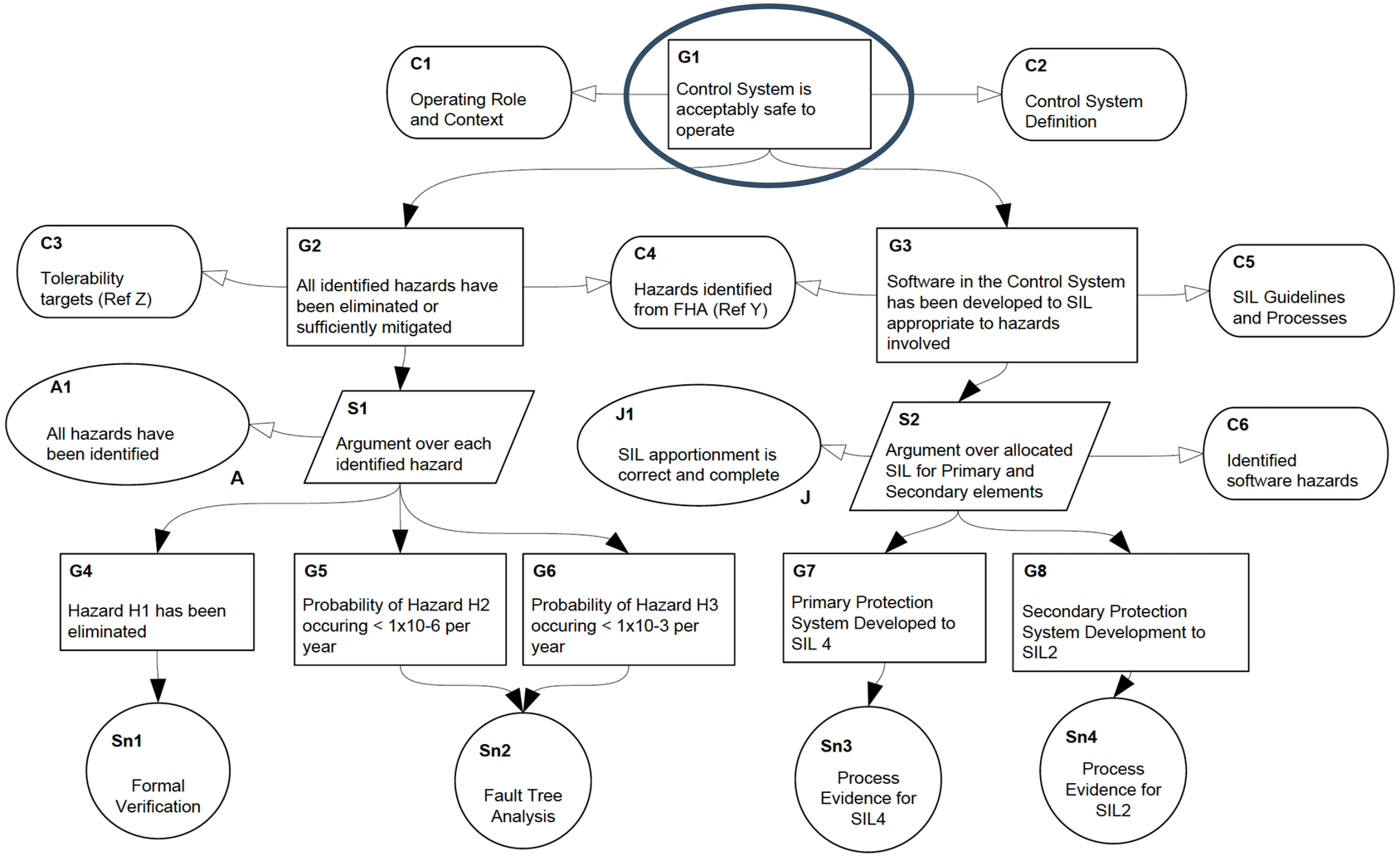
- Consider the basics of the argument-based assurance methodology
- Understand when and how it can be used to facilitate responsible communication
- Use the method to identify broader normative goals that may not have been covered in this course, and determine which properties need to be assured to help demonstrate that the respective goal has been obtained
- Reflect on what you have learned during the course

Day 5

What is Argument-Based Assurance?



“Argument-based assurance is a process of using structured argumentation to provide assurance to another party (or parties) that a particular claim (or set of related claims) about a property of a system is warranted given the available evidence.”



Safety Assurance

Assurance Cases

Assurance cases are a primary means by which confidence in the safety of the system is communicated to and scrutinised by the diverse stakeholders, including regulators and policy-makers.

Benefits of Assurance

Beyond Compliance

1. Assist internal reflection and deliberation by providing a systematic and structured means for evaluating how the development of systems or products can fulfil certain normative goals (e.g. safety or robustness), according to certain well-defined properties and criteria (e.g. risk reduction thresholds met)
2. Provide a deliberate means for the anticipation and pre-emption of potential risks and adverse impacts through mechanisms of end-to-end assessment and redress.
3. Facilitate transparent communication between developers and affected stakeholders.
4. Support mechanisms and processes of documentation (or, reporting) to ensure accountability (e.g. audits, compliance).
5. Build trust and confidence by promoting the adoption of best practices (e.g. standards for warranted evidence) and by conveying the integration of these into design, development, and deployment lifecycles to impacted stakeholders.

Preprint

Ethical Assurance

Full preprint is available 👉

https://arxiv.org/abs/2110.05164

Ethical Assurance

A practical approach to the responsible design, development, and deployment of data-driven technologies

Christopher Burr* David Leslie†

October 11, 2021

Abstract

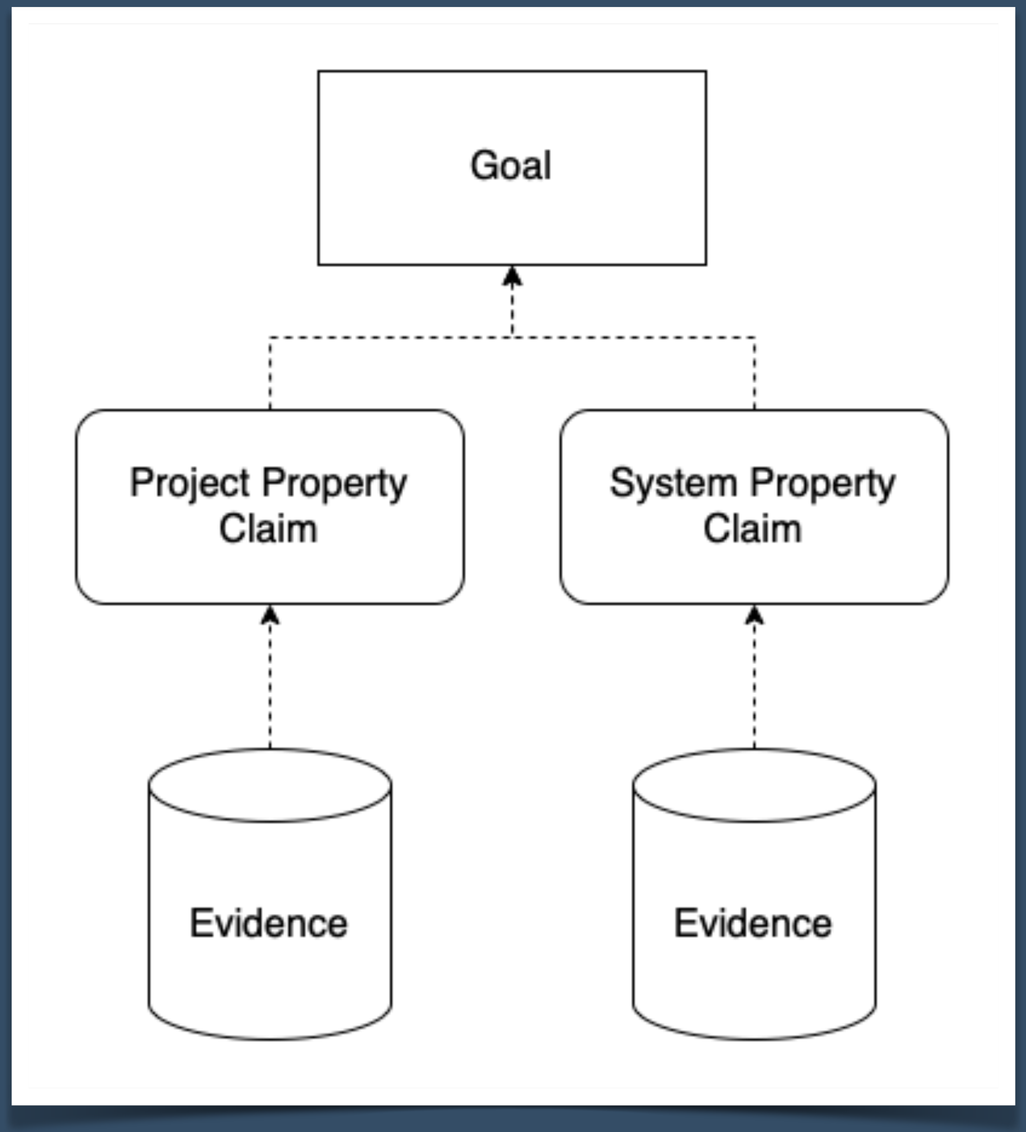
This article offers several contributions to the interdisciplinary project of responsible research and innovation in data science and AI. First, it provides a critical analysis of current efforts to establish practical mechanisms for algorithmic assessment, which are used to operationalise normative principles, such as sustainability, accountability, transparency, fairness, and explainability, in order to identify limitations and gaps with the current approaches. Second, it provides an accessible introduction to the methodology of argument-based assurance, and explores how it is currently being applied in the development of safety cases for autonomous and intelligent systems. Third, it generalises this method to incorporate wider ethical, social, and legal considerations, in turn establishing a novel version of argument-based assurance that we call 'ethical assurance.' Ethical assurance is presented as a structured means for unifying the myriad practical mechanisms that have been proposed, as it is built upon a process-based form of project governance that supports inclusive and participatory ethical deliberation while also remaining grounded in social and technical realities. Finally, it sets an agenda for ethical assurance, by detailing current challenges, open questions, and next steps, which serve as a springboard to build an active (and interdisciplinary) research programme as well as contribute to ongoing discussions in policy and governance.

1 Introduction

1.1 Setting the Stage

The recent history of artificial intelligence (AI) ethics and governance has been characterised by increasingly vocal calls for a move from *principles to practice*. Over the past several years, some have discerned a rapid transition in the

*The Alan Turing Institute, UK | Corresponding author: cburr@turing.ac.uk
†The Alan Turing Institute, UK

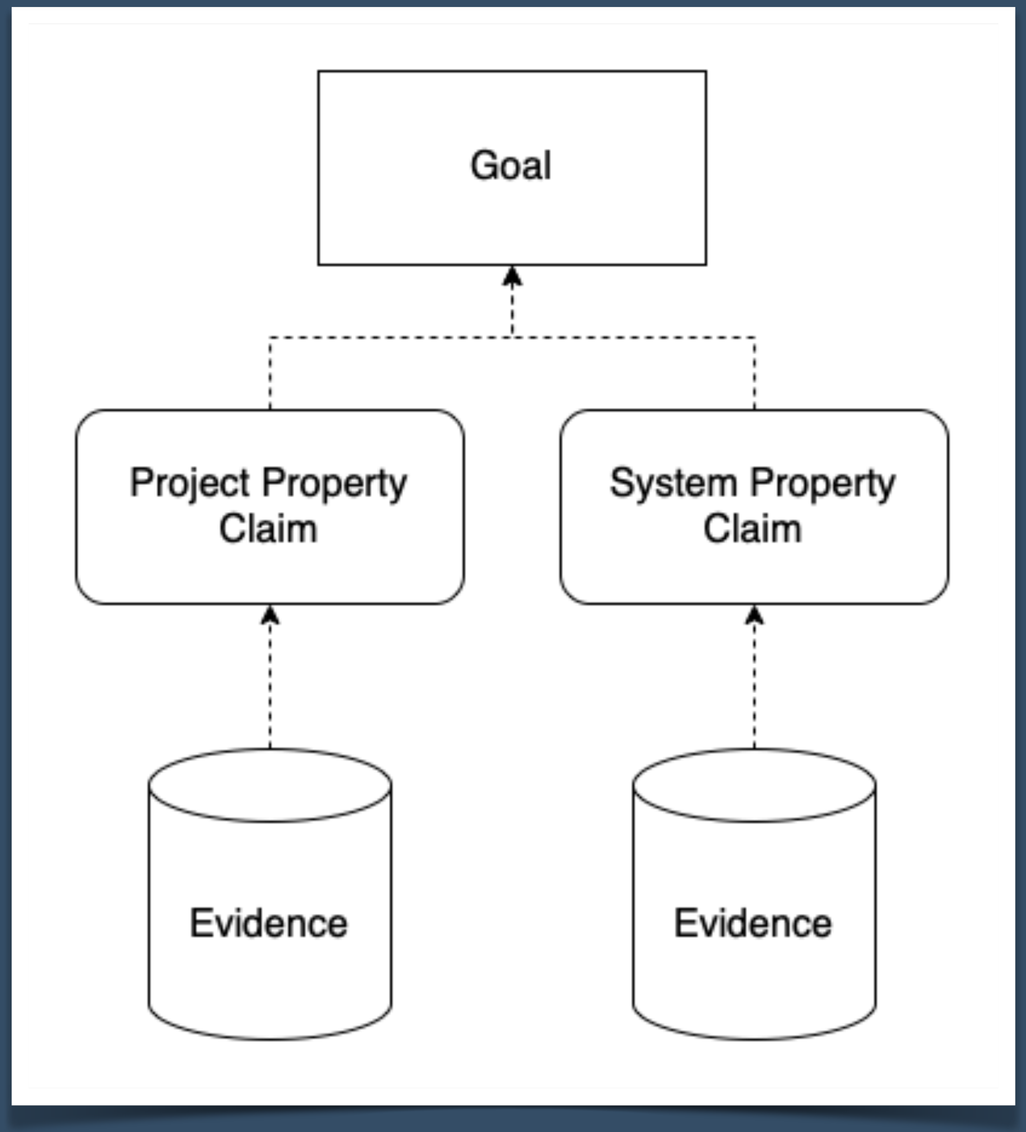


Elements and Structure

Building an Assurance Case

An ethical assurance case is built of the following (primary) elements:

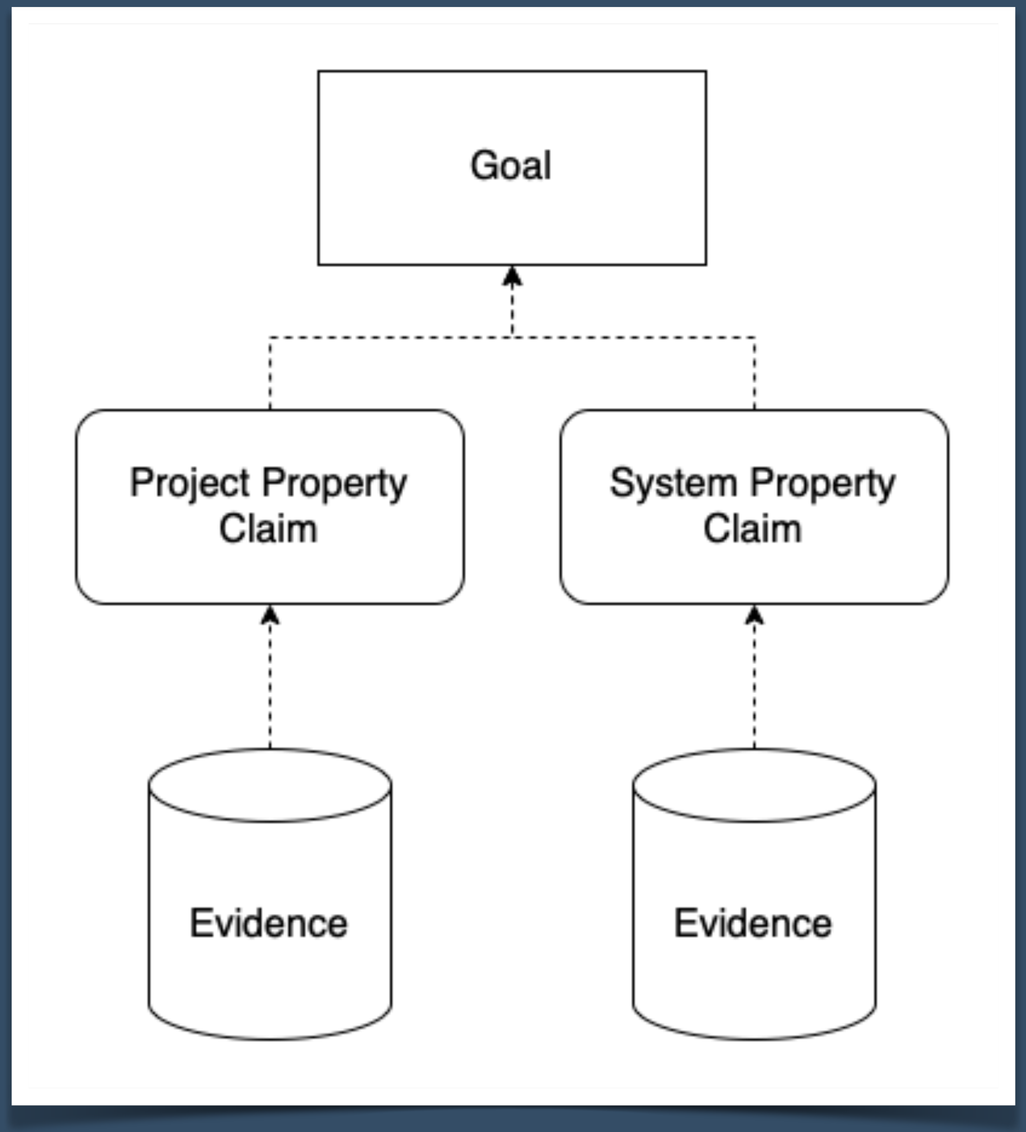
- Top-Level Ethical Goal
- (Set of) Project or System Property Claims
- Evidence (and Evidential Claims)



Elements and Structure

Building an Assurance Case

The lower level elements provide justificatory support for the higher-level elements in an abductive manner (i.e., defeasible but probable).



Elements and Structure

Building an Assurance Case

We can simplify the method used for safety cases into three steps:

- Reflect
- Act
- Justify

Reflect

Identify goals & challenges and determine how best to address them

Act

Take action(s) to address challenge and document steps taken at key stages of the project

Justify

Show how a specific claim about a project has been addressed, and why actions taken are warranted

Reflect

Identify goals & challenges and determine how best to address them

Reflection is an anticipatory and deliberative process in which questions such as the following are asked of the project and its governance:

1. What are the goals of your system?
2. How are these goals defined?
3. Which stakeholders have participated in the identification and defining of these goals?
4. What properties need to be implemented in the project or system to ensure that these goals are achieved?
5. Which actions ought to be taken to establish these properties within the project or system?

Reflect

Identify goals & challenges and determine how best to address them

Act

Take action(s) to address challenge and document steps taken at key stages of the project

Justify

Show how a specific claim about a project has been addressed, and why actions taken are warranted

Act

Take action(s) to address challenge and document steps taken at key stages of the project

Action occurs throughout all of the stages of the project lifecycle, and the output of many of these actions are likely to serve as the evidence for the claims of the assurance case. These actions and evidential artefacts can also help you identify what claims may be relevant in your argument:

1. What actions have been undertaken during **(project) design** that have generated salient evidence for your goals and claims?
2. What actions have been undertaken during **(model) development** that have generated salient evidence for your goals and claims?
3. What actions have been undertaken during **(system) deployment** that have generated salient evidence for your goals and claims?

Reflect

Identify goals & challenges and determine how best to address them

Act

Take action(s) to address challenge and document steps taken at key stages of the project

Justify

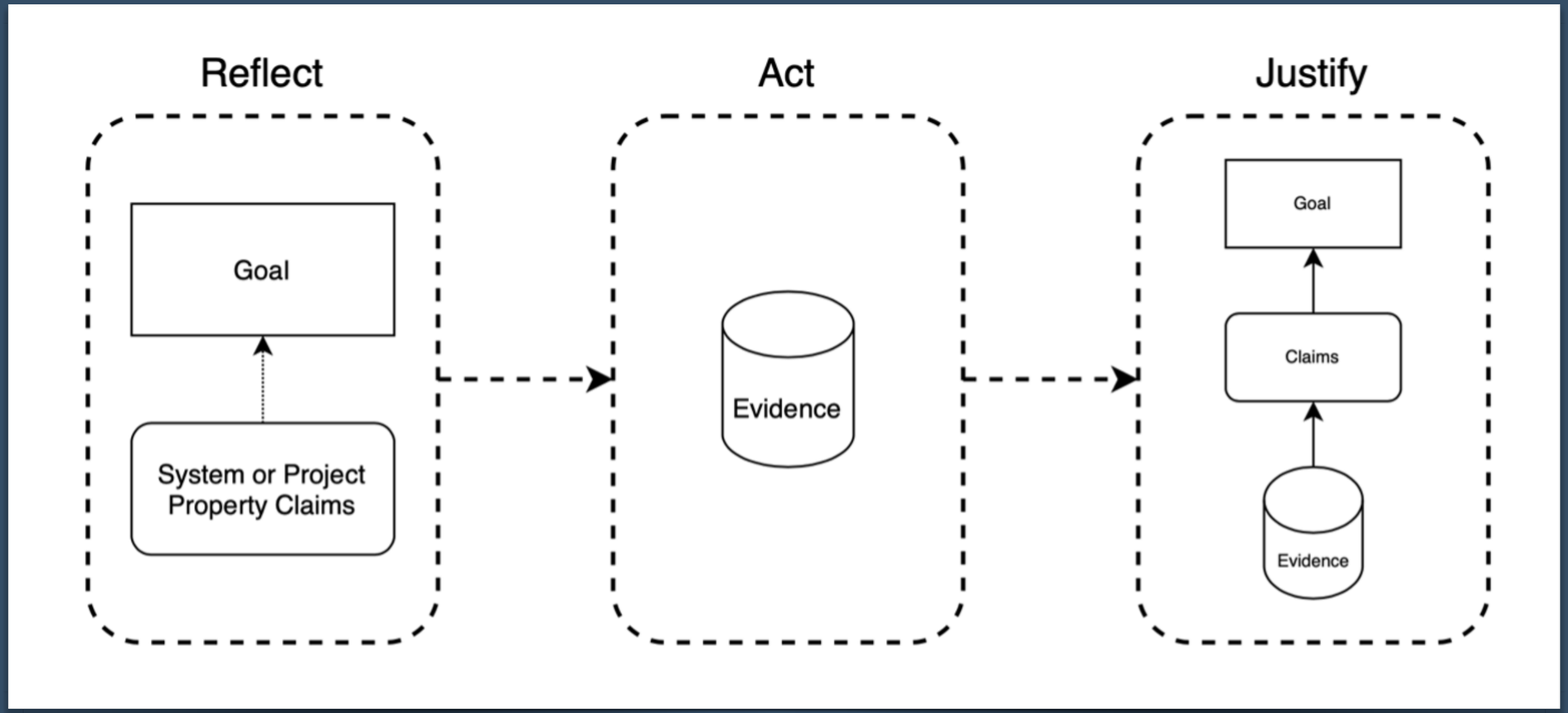
Show how a specific claim about a project has been addressed, and why actions taken are warranted

Justify

Show how a specific claim about a project has been addressed, and why actions taken are warranted

The final step is to **justify** that your evidence base is sufficient to warrant the claims that are being made about the properties of your project or system. This does not mean that the assurance case is the final activity that needs to be done at the very end of a project. Rather, it’s development should be seen as iterative and ongoing as the project evolves.

1. Which stakeholders, identified in your stakeholder engagement plan, can support the evaluation of your evidence and overall case?
2. Is any evidence missing from your case?
3. Are the collection of property claims jointly sufficient to support your top-level goal?

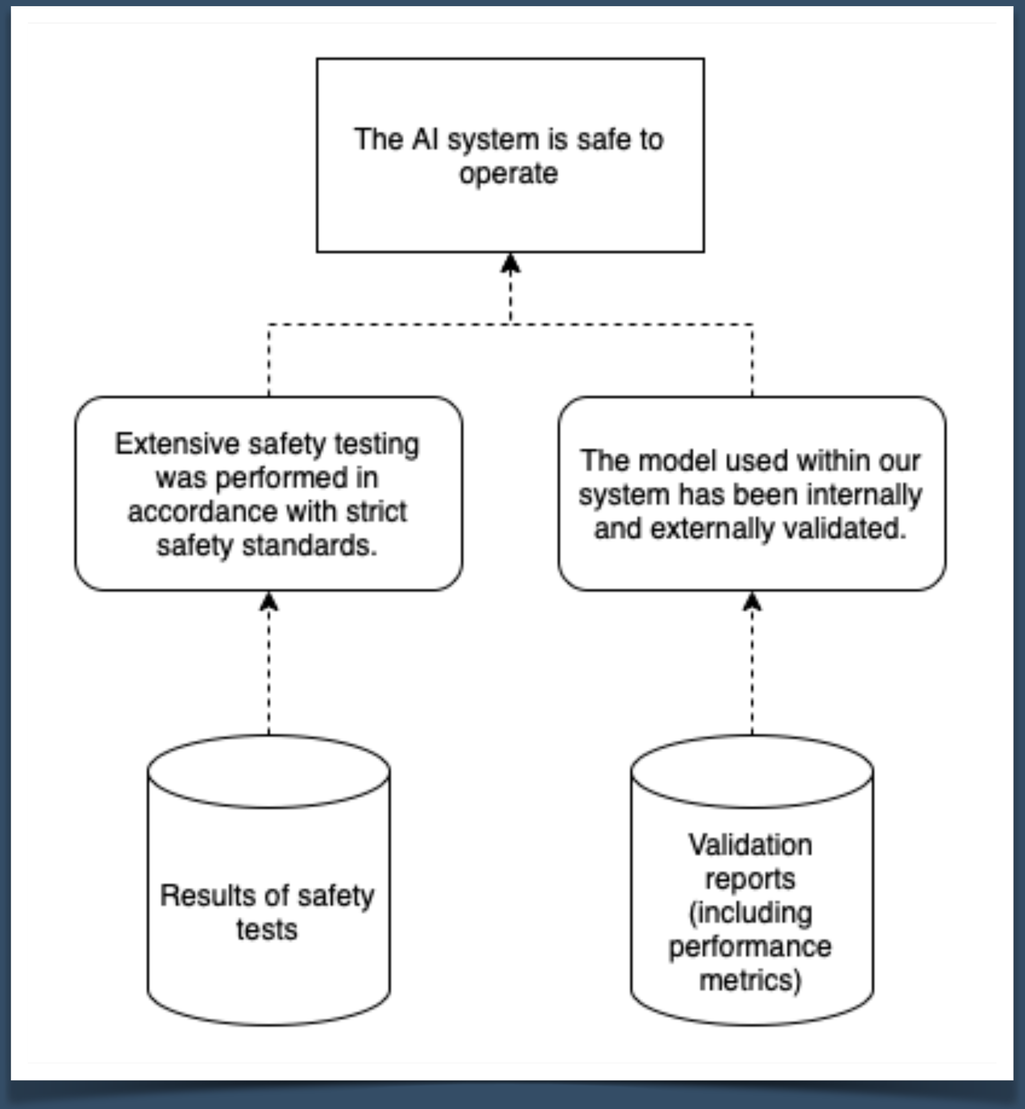

Reflect
Goal
System or Project Property Claims
Act
Evidence
Justify
Goal
Claims
Evidence

Example 1

Sustainability Goal

Here is an example of an (incomplete) assurance case that is directed towards a safety-based conception of the sustainability goal.

Note the properties of the claims that are referred to.

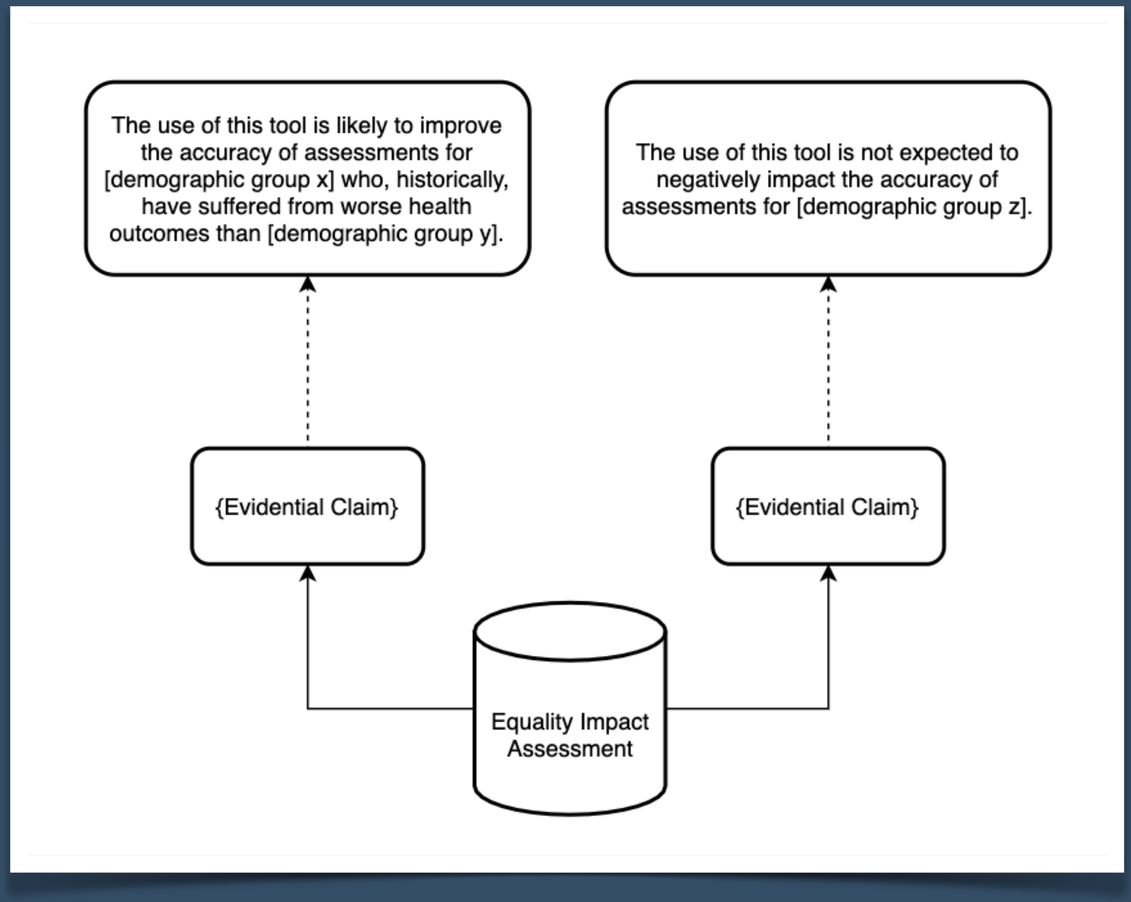


Example 2

Fairness Evidence

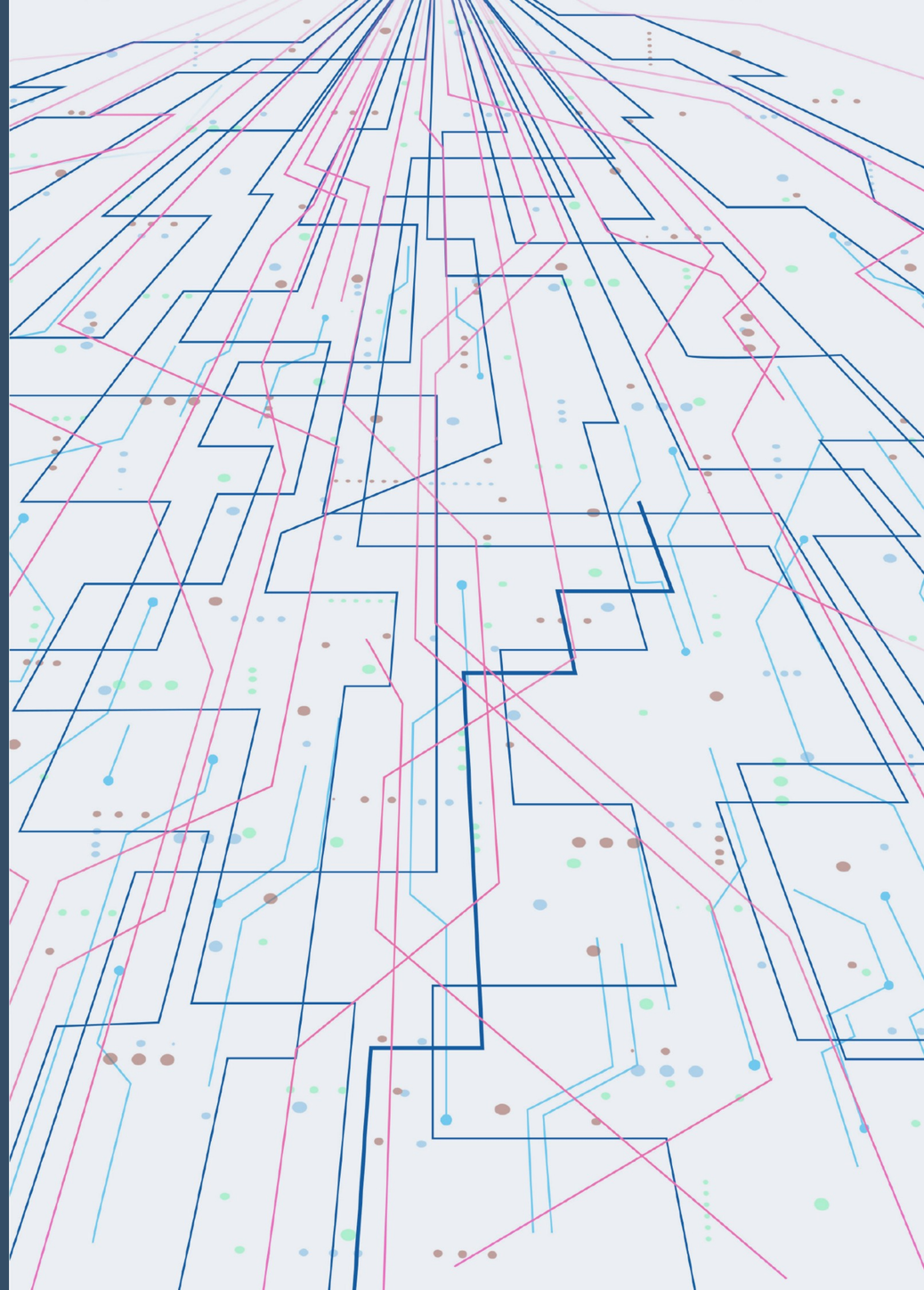
Here is another example focusing on the evidential artefact that can be used to support two separate claims.

In this instance, two separate evidential claims are required to direct readers to the relevant part of the document (i.e., an equality impact assessment).



Day 5

Goals, Properties, and Evidence



SAFE-D Principles

Sustainability

Sustainability requires the outputs of a project to be:

- safe, secure, robust, and reliable
- informed by ongoing consideration of the risk of exposing individuals to harms even after the system has been deployed and the project completed—a long-term (or sustainable) form of safety.

Core Attributes

Sustainability

- Safety
- Security
- Robustness
- Reliability
- Accuracy and Performance

SAFE-D Principles

Accountability

Accountability can refer to transparency of processes and associated outcomes that enable people to understand how a project was conducted (e.g., project documentation), or why a specific decision was reached. But it can also refer to broader processes of responsible project governance that seek to establish clear roles of responsibility where full transparency may be inappropriate (e.g., confidential projects).

Core Attributes

Accountability

- Traceability
- Answerability
- Auditability
- Clear Data Provenance and Lineage
- Accessibility
- Reproducibility

SAFE-D Principles

Fairness

Fairness is inseparably connected with legal conceptions of equality and justice, which may emphasise a variety of features such as non-discrimination, equitable outcomes, or procedural fairness through bias mitigation.

However, these notions serve as a subset of broader normative considerations pertaining to social justice, socioeconomic capabilities, diversity and inclusivity.

Core Attributes

Fairness

- Bias Mitigation
- Diversity and Inclusiveness
- Non-Discrimination
- Equality

SAFE-D Principles

Explainability

Explainability is a key condition for autonomous and informed decision-making in situations where data-driven systems interact with or influence human judgement and choice behaviour.

Explainability goes beyond the ability to merely interpret specific aspects of a project (e.g., interpreting the parameters of a model); it also depends on the ability to provide an accessible and relevant information base about the processes behind the outcome.

Core Attributes

Explainability

- Interpretability
- Responsible Model Selection
- Accessible Rationale Explanation
- Implementation and User Training

SAFE-D Principles

Data Quality

‘Data Quality’ captures the static properties of data, such as whether they are (a) relevant to and representative of the domain and use context, (b) balanced and complete in terms of how well the dataset represents the underlying data generating process, and (c) up-to-date and accurate as required by the project.

Core Attributes

Data Quality

- Source Integrity and Measurement Accuracy
- Timeliness and Recency
- Relevance, Appropriateness, and Domain Knowledge
- Adequacy of Quantity and Quality
- Balance and Representativeness

SAFE-D Principles

Data Integrity

‘Data Integrity’ refers to more dynamic properties of data stewardship, such as how a dataset evolves over the course of a project lifecycle. In this manner, data integrity requires (a) contemporaneous and attributable records from the start of a project (e.g., process logs; research statements), (b) ensuring consistent and verifiable means of data analysis or processing during development, and (c) taking steps to establish findable, accessible, interoperable, and reusable records towards the end of a project’s lifecycle.

Core Attributes

Data Integrity

- Attributable
- Consistent, Legible and Accurate
- Complete
- Contemporaneous
- Responsible Data Management
- Data Traceability and Auditability

SAFE-D Principles

Data Protection and Privacy

‘Data protection and privacy’ reflect ongoing developments and priorities as set out in relevant legislation and regulation of data practices as they pertain to fundamental rights and freedoms, democracy, and the rule of law. For example, the right for data subjects to have inaccurate personal data rectified or erased.

Core Attributes

Data Protection and Privacy

- Consent (or legitimate basis) for processing
- Data Security
- Data Minimisation
- Transparency
- Proportionality
- Purpose Limitation

Examples

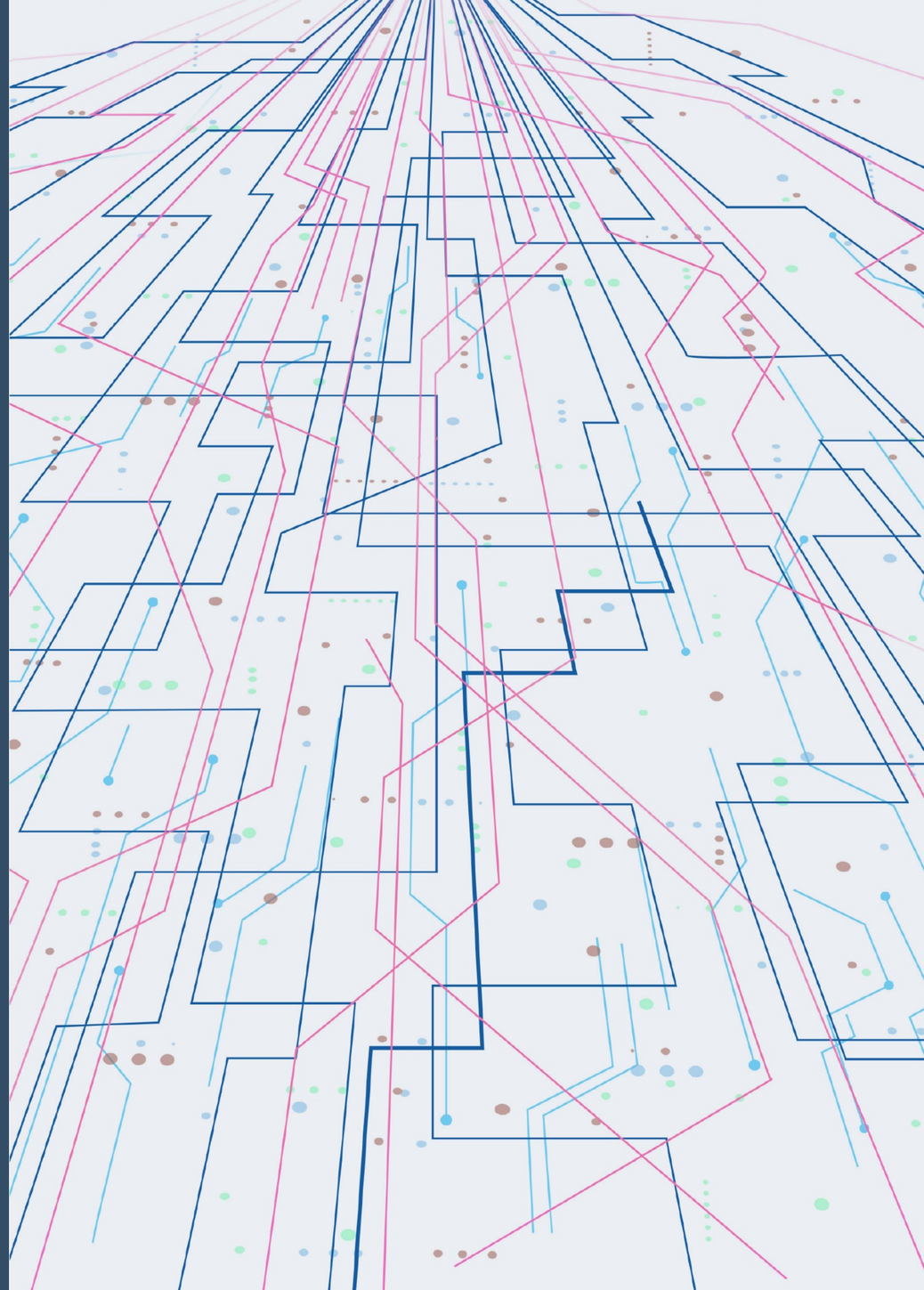
Property Claims

The following claims offer examples for each of the SAFE-D principles (or, goals) along with an attribute and corresponding project lifecycle stage

Goal & Attribute	Example Property Claim	Project Lifecycle Stage
Sustainability (Robustness)	*The model used in our system has been internally and externally validated. The external validation has been carried out across several varied environments to ensure robustness of the system.*	Model Training, Testing and Validation
Accountability (Accessibility)	*All identified stakeholders were consulted prior to the development of our system to help critically evaluate our project plans and ensure they were intelligible.*	Project Planning and Problem Formulation
Fairness (Equality)	*Persons affected by use of the system have avenues of recourse, ability to contest system outputs and demand human intervention.*	System Use & Monitoring
Explainability (Responsible Model Selection)	*Features were hand-selected in conjunction with domain experts to optimise for both interpretability and predictive power.*	Preprocessing & Feature Engineering and Model Selection
Data Quality (Timeliness & Recency)	*Only data that were collected within the previous 3 months were used to ensure the training data were up-to-date.*	Data Extraction or Procurement

Activity 10

Building an Assurance Case



Breakout Groups

Plenary

Next Steps

Wrap-Up

- Final Feedback Form
- Guidebook (Online & PDF)
- Further Resources
- GitHub Contributions
- Future Courses

Thank you!

I hope you've enjoyed this course.

The Alan Turing Institute

Dr Christopher Burr
cburr@turing.ac.uk
@chrisdburr