

The Alan Turing Institute

Trustworthy and Ethical Assurance of Digital Health and Healthcare

Supporting an assurance
ecosystem for data-driven
technologies in health
and healthcare



This report is produced as part of the Trustworthy and Ethical Assurance of Digital Health and Healthcare (TEA-DH) project, which was generously funded by the Assuring Autonomy International Programme, a partnership between Lloyd's Register Foundation and the University of York.

We would like to thank all the workshop participants that provided support and anonymous feedback as part of this project's engagement activities.

We would also like to thank the following reviewers, who kindly offered their feedback on the draft version of this report: Lara Groves (Ada Lovelace Institute), Mahi Hardalupas (Ada Lovelace Institute), Chris Thomas (The Alan Turing Institute), Mark Sujun (University of Warwick).

When citing this version of the report, please use the following details:

Burr, C., Arana, S., Gould Van Praag, C., Habli, I., Kaas, M., Katell, M., Laher, S., Leslie, D., Niederer, S., Ozturk, B., Polo, N., Porter, Z., Ryan, P., Sharan, M., Solis Lemus, J. A., Strocchi, M., Westerling, K., (2024)

Trustworthy and Ethical Assurance of Digital Health and Healthcare.

<https://doi.org/10.5281/zenodo.10532573>

**The
Alan Turing
Institute**



Responsible
Technology
Adoption Unit



UNIVERSITY
of York

Table of Contents

Executive Summary	4
-------------------	---

Chapter	Introduction	> Health and Healthcare	12
1		> What is Trustworthy and Ethical Assurance?	15
		> From Safety to SAFE-D: Operationalising Ethical Principles	22

Chapter	Fairness and Health Equity: The Unvirtuous Circle	> Social Determinants of Health	30
2		> Two Models for Supporting Fairness Assurance in Digital Health and Healthcare	32

Chapter	A Plurality of Approaches: Assuring Fairness and Health Equity	> Case Study 1: AI-enabled clinical diagnostic support system	44
3		> Case Study 2: Cardiac Electro-Mechanics Research Group Application	51

Chapter	Scaffolding Communities of Practice	> What is a Community of Practice?	60
4		> Addressing Bias in Assurance Cases: The Case for Public Reason	64
		> Next Steps: Making Fair Assurance Cases FAIRer	64

References	71
------------	----

Appendix 1: Trustworthy and Ethical Assurance of Digital Health and Healthcare (TEA-DH) Project and Engagement	78
--	----

Appendix 2: The Trustworthy and Ethical Assurance (TEA) Platform	84
--	----

Executive Summary

As data-driven technologies, such as digital twins or AI systems, continue to be used in critical sectors like healthcare, finance, and criminal justice, ensuring they are designed, developed, and deployed in a trustworthy and ethical manner is paramount. Unchecked biases, opaque algorithms, and security vulnerabilities are not mere theoretical hazards; they can lead to real-world harms, including perpetuating and exacerbating existing discrimination, amplifying societal inequities, and even endangering lives. In this context, neglecting to assure that data-driven technologies are trustworthy and ethical isn't just a technical oversight, it's a critical moral failure with potentially severe consequences for individuals and society. To harness the full potential of data-driven technologies while mitigating the inherent risks, we must prioritise building systems that are trustworthy and ethical. People will not adopt or use systems that are inherently unfair or untrustworthy.

This report introduces a technique for assurance of data-driven technologies, known as *Trustworthy and Ethical Assurance* (or TEA). In general, assurance helps build trust and confidence in a system, product, or process by measuring, evaluating, and communicating relevant properties. More specifically, TEA is an example of *argument-based assurance*, which emphasises the importance of developing and communicating an *accessible and structured argument* that helps affected stakeholders understand the claims and evidence that justify confidence in a related goal (e.g. developing a fair system).

There are several reasons why TEA can be valuable in the context of the broader assurance ecosystem:

-
1. It extends traditional frameworks of assurance, which conventionally focus on principles such as safety, to consider wider ethical goals, such as fairness.

 2. It establishes a systematic method for specifying and operationalising ethical principles and normative goals, such as fairness or explainability.

 3. It allows people to query and evaluate the strength of evidence that is used to justify the validity of specific claims and the overarching argument about the ethics and trustworthiness of an AI system, promoting a more open, dialogical, and collaborative approach to assurance.

 4. It facilitates the development of *communities of practice*, which can use the framework as a means for identifying best practices or addressing gaps in the current regulatory and governance ecosystem.

While there are good reasons to believe that these benefits will apply broadly [1], this report discusses them in the specific context of digital health and healthcare. However, rather than focusing on assurance goals such as *clinical efficacy and safety*, which have been explored by others [1], [2], we consider the specific goals of *fairness and equity*.

Ensuring fair and equitable access to healthcare has always been an important goal. However, there are two reasons why this goal is even more pressing today. First, the COVID-19 pandemic exposed and exacerbated weaknesses and vulnerabilities in our healthcare systems, widening the gap between the most vulnerable and the most advantaged, and fuelling higher infection and mortality rates in marginalised communities [3]. Second, and as mentioned above, unchecked biases in data-driven technologies or systems can replicate, exacerbate, and perpetuate existing societal inequities, which is why so many have already drawn attention to such risks in the literature [4], [5], [6], [7], [8], [9]. This attention to principles of fairness and equity is reflected in recent regulatory guidance, such as the recent policy paper, 'A pro-innovation approach to AI regulation', produced by the UK's Department for Science, Innovation and Technology, and AI Policy Directorate [10].

What are the primary goals of this report?

1. To introduce and motivate the TEA platform by demonstrating its utility within the domain of digital health and healthcare.
2. To contextualise this work by situating it within some of the existing and emerging AI policy and governance developments.
3. To demonstrate how the TEA platform can be used to assure ethical principles, using two case studies focused on the goal of 'fairness and health equity'.
4. To establish an agenda and set of next steps for the TEA platform, grounded in a community-centred approach.



Key Resources

For those who prefer to get hands-on with the TEA platform, you can access the documentation and open-source tool by going straight to our documentation site or GitHub repository.

- › Documentation and User Guidance: <https://alan-turing-institute.github.io/AssurancePlatform>
- › GitHub repository: <https://github.com/alan-turing-institute/AssurancePlatform>

Who is this report for?

This report is wide-ranging in its content and scope. As such, there is no single audience. Rather, in line with the recent framework and guidance from the UK's Department for Science, Innovation, and Technology (Responsible Technology Unit) this report has been written to be accessible to various actors within the assurance ecosystem [11]. This includes:

- › Regulators (e.g. Medicines and Healthcare products Regulatory Agency; Information Commissioner's Office; US Foods and Drugs Administration; European Medicines Agency)
- › Accreditation bodies (e.g. United Kingdom Accreditation Service)
- › Government
- › Standards Bodies (e.g. ISO, the International Organization for Standardization; British Standards Institution)
- › Research Bodies (e.g. universities)
- › Civil Society Organisations (e.g. Ada Lovelace Institute)
- › Professional Bodies (e.g. International Association of Algorithmic Auditors)
- › Other assurance providers
- › Affected Individuals

However, we also include *practitioners* more generally as key contributors to the assurance ecosystem (e.g. data science professionals¹, software engineers, product managers).²

The decision to have such a broad scope comes with an unavoidable trade-off—it sacrifices more targeted and actionable recommendations or research findings in favour of breadth. However, the choice is intentional and in line with a key ambition of the TEA platform, which is to create a more inclusive and accessible set of assurance tools and mechanisms that enable greater participation in the assurance ecosystem.

¹ See <https://alliancefordatascienceprofessionals.co.uk> for further information.

² This group would be best captured as 'Organisations developing AI systems' and 'Organisations procuring AI systems' within [11], but for our purpose is too specific as we are not solely concerned with AI systems.

How should this report be read?

Because of the above decision, the following table can help readers determine which sections are most relevant to them:

Sections and intended audiences in the report		
SECTION	DESCRIPTION	AUDIENCE
1	<p>We set the context for Trustworthy and Ethical Assurance against the backdrop of current and emerging regulatory approaches to AI risk identification, management, and mitigation. This includes a summary of the AI Assurance Ecosystem, as developed by the UK's Department for Science, Innovation, and Technology.</p> <p>Following this, we introduce the TEA framework.</p>	All (and, specifically, any audience unfamiliar with argument-based assurance or TEA)
2	<p>We delve deeper into the topics of fairness and equity in health and healthcare, discussing relevant issues such as the <i>social determinants of health</i>. Building on previous work, we introduce two models for identifying and mitigating fairness-related risks in the design, development, and deployment of digital health and healthcare technologies.</p>	Research Bodies Practitioners Civil Society Organisations Standards Bodies Affected Individuals

SECTION	DESCRIPTION	AUDIENCE
3	<p>We present two case studies to help illustrate how fairness can be assured over the course of a project's lifecycle. The first project involves an AI-enabled clinical diagnostic support system (CDSS) the purpose of which is to help clinicians predict the risk of developing hypertension in patients with Type 2 diabetes. The second presents a medical imaging platform, known as the CemrgApp, with custom image processing and computer vision toolkits for applying statistical, machine learning, and simulation approaches to cardiovascular data.</p> <p>Both case studies provide illustrative examples of different approaches to assuring fairness and health equity, using the TEA platform and associated materials. As such, they demonstrate the practical utility of the platform.</p>	<p>Research Bodies</p> <p>Practitioners</p> <p>Regulators</p> <p>Accreditation Bodies</p> <p>Civil Society Organisations</p> <p>Standards Bodies</p> <p>Professional Bodies</p>
4	<p>We conclude by discussing how TEA can be enhanced and extended by supporting the development of <i>communities of practice</i>. We introduce what is meant by this term, why it matters for assurance of digital health and healthcare, and how we plan to build on this work to ensure the sustainable development of the TEA framework.</p>	<p>Research Bodies</p> <p>Practitioners</p> <p>Regulators</p> <p>Affected Individuals</p> <p>Civil Society Organisations</p> <p>Standards Bodies</p>
Appendices	<p>Two appendices provide supplementary information:</p> <p>Appendix 1 presents information about the Trustworthy and Ethical Assurance of Digital Health and Healthcare (TEA-DH) project, which gave rise to this report. This includes information about a series of engagement workshops that helped shape the development of the TEA platform and the guidance offered in this report.</p> <p>Appendix 2 presents information about the TEA platform, including how to access it and the features available or planned for future development.</p>	<p>All (based on need for additional information specified throughout the report)</p>

01



Introduction

Health and Healthcare

What is Trustworthy and Ethical Assurance?

› Methodology: Argument-Based Assurance

From Safety to SAFE-D: Operationalising Ethical Principles

By now, it has been well established that the use of data-driven technologies, such as machine learning (ML), artificial intelligence (AI), or digital twins (DTs), changes the risk profile of organisations, including those in health and healthcare. The identification, evaluation, and management of risks and benefits for any novel technology can be challenging. But sociotechnical systems³ that include forms of AI pose their own unique difficulties, and as such have received a lot of regulatory attention and scrutiny.

Over the last several years, many national and multi-national science, innovation, and technology governance bodies and organisations have worked tirelessly to develop their respective visions of how to manage and govern the risks associated with AI systems in a responsible and trustworthy manner⁴. For instance, the European Commission's AI regulatory framework [12] adopts an approach in which systems are classified according to levels of risk (alongside dedicated rules for general purpose AI [13]), with subsequent rules that apply conditionally upon the initial risk classification (see **Figure 1.1**). The latest version of the AI Act also includes a fifth level for General-Purpose AI Models with "systemic risk" (see Article 52d [14]).

Similarly, the US National Institute of Standards and Technology published the first version of their AI Risk Management Framework (AI RMF) at the beginning of 2023 [15], which is designed to help users map, measure, and manage risks (see **Figure 1.2**), while also exploring the broader trustworthiness of AI as dependent upon related properties, such as 'security', 'explainability', and 'fairness' (see **Figure 1.3**).

Although there are important differences between the two frameworks, such as the voluntary nature of the latter, they both share a commitment to a *risk-based approach*. This is important, because there are, obviously, many different harms that can emerge during the design, development, and deployment of data-driven technologies such as AI. As such, being able to categorise risks can help organisations adopt a proportionate approach to managing them.

However, there are also (well-known) challenges and limitations with risk-based approaches, including:

- › **Missed hazards:** subjectivity in risk assessments can lead to some hazards being overlooked whether unintentionally or perhaps due to variation in risk thresholds.
- › **Dynamic and evolving risks:** the behaviour of data-driven technologies can change over time (even in deployment where they utilise forms of AI). As such, new risks may emerge as systems adapt or evolve. Risks assessments may not capture these emerging hazards.

³ In short, systems where people, technology, social environments, and work processes influence each other.

⁴ While we focus primarily on AI in the context of this section, the scope of our report and the TEA platform is broader than AI. The emphasis on AI is a product of the regulatory and governance frameworks we explore for the purpose of scene-setting and context. However, in the context of assurance more generally, the scope of our concern is better captured by the term 'data-driven technology', which we define as any technology that leverages data to learn, adapt, and function, including AI, machine learning, and digital twins. Such a definition would be too broad and inclusive for the purpose of governance and regulation but is suitable for our purposes.



Figure 1.1: Two graphics that (a) depict the four risk categories of the EU’s Regulatory Framework for AI systems, and (b) outline the critical steps for high-risk AI systems to meet EU compliance standards before reaching the marketplace (reprinted from [12]).



Figure 1.2: NIST’s AI Risk Management Framework encapsulates the essential stages of Mapping, Measuring, Managing, and Governing AI risks.



Figure 1.3: Characteristics of trustworthy AI systems, as described by NIST’s AI RMF.

- › **Availability of data:** effective risk assessments depend on the quality and relevance of the available data. Where data are incomplete or biased, risk assessments may lead to ineffective governance strategies.
- › **Under-regulation and governance of low-risk systems:** low-risk doesn't mean no risk. It is common for risk management and mitigation to be proportional to the risk. However, this means that low risk systems may not be given sufficient attention, which could lead to low risks becoming high risks where systems are combined (e.g. in procurement) or when non-independent risks are aggregated (e.g. pooling of sub-prime mortgages in 2008 financial crash).
- › **Limited user or stakeholder engagement:** in some cases, stakeholders, end-users, or the individuals impacted by data-driven technologies may not be adequately involved in the risk assessment process. Rather, it may be treated as a mere *compliance* exercise, to be completed by an expert. This contributes to the first issue (i.e. missed hazards), and also misses the opportunity to build social license and capacity within the relevant community or ecosystem.

This is not a complete list (see [16] for further discussion), but helps to identify some of the reasons why we argue in favour of a more comprehensive and open approach to process-based of governance and assurance, which we will discuss in the context of health and healthcare.

Health and Healthcare

In health and healthcare, possible benefits from data-driven technologies include improved diagnostic accuracy and clinical decision support, earlier prediction of disease and associated health outcomes, and operational efficiencies in healthcare pathways, such as using AI to forecast how much blood plasma a hospital needs to hold onsite on any given day [17].

! Caution:

- Before activating Production use for MedLM, customers must reach out to Google Product Team to discuss usage.
- MedLM has not been designed or developed to be used as a medical device. Any output should be verified by a Healthcare Professional (HCP), and no direct diagnosis should be claimed.
- The generated output may not always be completely reliable. Due to the nature of LLMs and Generative AI, outputs may have incorrect or biased (for example, stereotypes or other harmful content) information and should be reviewed. All summaries or answers should be considered draft and not final.
- If Vertex AI detects content that violates our policies, including [Google Cloud Platform Acceptable Use Policy](#) and [Generative AI Prohibited Use Policy](#), a response is not returned.
- When used by HCPs for Q&A purposes, MedLM is only intended for use as an educational tool for medical training or to reinforce the HCP's prior training.
- LLM output may not follow the exact format laid out in the prompt. The prompt design to extract information for each field should take into account that the format may deviate from the original (for example, dashes in field names, exact capitalization of letters).

Figure 1.4: A list of cautions that accompany Google's MedLM foundation model. Accessed 02/01/2024 from <https://cloud.google.com/vertex-ai/docs/generative-ai/medlm/overview>

On the other hand, hazards (and their associated risks)⁶ include encoded bias in datasets, opaque and uninterpretable models, increased bureaucratic overhead to operate or support technical systems, and deskilling of professional judgement due to overreliance on algorithmic, automated, or autonomous systems. Some of these benefits and hazards are already being realised or occurring, causing issues for accountability, liability, and justifiability of decisions [18], [19], whereas others remain possibilities. And in many cases, the benefits and hazards are interdependent.

For instance, the use of generative AI, such as large-language models, has been touted by many as a promising source of clinical value. However, as healthcare is often a high-risk environment, any potential benefit must be carefully weighed against the possible risks in a proportional manner and relative to one other. As such, even models that are fine-tuned for the healthcare domain, such as Google’s MedLM are accompanied by wide-ranging cautionary remarks (see **Figure 1.4**).

However, it is insufficient to just note that ‘benefits’ and ‘hazards’ of such technologies exist. And although risk management frameworks, as noted previously, are helpful for offering top-down guidance, they can fail to provide adequate forms of specification for how ethical principles should be applied and how specific risks should be handled and communicated. Risks still need to be properly specified, evaluated, and mitigated (e.g. identifying hazards to physical or psychological harm and safety processes implemented to mitigate them), and then subsequently this risk management process should be sufficiently documented and communicated to relevant stakeholders or affected users.



Figure 1.5: A summary of the AI Ethics and Governance in Practice model, showing how the process-based governance framework is grounded in ethical values and helps operationalise key ethical principles across a project’s lifecycle [22].

⁶ We can define ‘hazards’ as a source or situation with the potential to cause harm (e.g., a sharp knife, a wet floor, encoded bias). Hazards exist independently of whether someone is exposed to them. A ‘risk’ is the chance, likelihood, or probability that harm will occur from a hazard, and the severity of that harm.

The need for thorough, end-to-end documentation is one of the reasons why we argue in favour of a broader approach to assurance in this report. An approach that still encompasses tools and process for risk management and mitigation, but also focuses on the need to build trust and confidence in a system by measuring, evaluating, communicating properties and evidence that contribute to its trustworthiness and ethical permissibility. Such a process cannot be reduced to a single activity or exercise (e.g. a risk assessment or production of a model card [20]). Rather it requires ongoing reflection, deliberation, and engagement across the entire lifecycle of a product or system, as noted by many researchers in the space of AI ethics and governance [21], [22], [23], [24].

The UK's Public Sector Guidance on AI Ethics and Safety characterises this end-to-end form of process-based governance as a three-step process of *reflection*, *action*, and *justification* [25]. Since the publication of this guidance in 2019, we have worked to expand this guidance into a comprehensive set of tools, guidance, and standards that can help scaffold and build capabilities around the responsible and ethical governance of AI systems [22].

For instance, as **Figure 1.5** shows, rather than having one over-arching notion of risk, there are more specific ethical risks captured by the SSAFE-D principles (e.g. risks associated with fairness or explainability). These risks are then operationalised in the context of a process-based governance (PBG) framework, which shows how specific actions or tools (e.g. software packages for quantifying statistical fairness or model interpretability) ground risk management practices in specific stages of a project's lifecycle. The purpose of the PBG Framework is to help project teams to successfully operationalise ethical principles across the AI project lifecycle and document how this has been achieved. As such, it is a form of scaffolding or a template that provides a landscape view of where in the AI project workflow governance actions are to take place to integrate the respective principles into project activities.

And, importantly, the emphasis on justification and documentation of governance actions ensures that risk management becomes (where possible) an open and dialogical process. Or, in another sense, it turns risk management to a more comprehensive form of assurance (i.e. establishing justified trust and confidence in some property or system). As such, this framework is a grounding for our approach to assurance. So, let us take some time now for TEA!

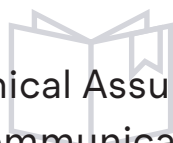


Further Resources

Rather than fully describing process-based governance here, we direct interested readers to a series of guidebooks designed to help the public sector apply AI ethics and safety to the design, development, and deployment of algorithmic systems:

› <https://www.turing.ac.uk/research/research-projects/ai-ethics-and-governance-practice>

What is Trustworthy and Ethical Assurance?



Trustworthy and Ethical Assurance is a structured approach to the communication of reasons and evidence about a data-driven technology or system, which helps stakeholders and affected users understand and evaluate the trustworthiness and validity of an argument made about some property or goal of the technology or system.

Trustworthy and ethical assurance comprises a *tool* and a *framework*.

The open-source tool, known as the TEA platform, has been designed and developed by researchers at the Alan Turing Institute and the University of York, with input from the Responsible Technology Adoption Unit, to support the process of providing assurance for some property or goal of a data-driven technology or system (e.g. AI system).

In addition, the TEA platform is supported and enhanced by a *framework* that offers accessible resources and user guidance to help scaffold a *community of practice*⁷ and shared standards. At the centre of the TEA framework are two components:

-
1. a *methodology* of argument-based assurance, focused on ethical goals, which provides the structure and elements necessary to build accessible assurance cases, and
 2. a set of *resources* that support practical decision-making across all stages of a project's lifecycle.
-

Here, we will focus on the methodology. Further details of both the tool and the framework can be found in **Appendix 2**.

Methodology: Argument-Based Assurance

Trustworthy and Ethical Assurance is an example of what is known as *argument-based assurance*. As we discuss in **Box 1.1: The AI Assurance Ecosystem**, there are many ways of providing assurance. However, TEA emphasises the importance of an *accessible and structured*

⁷ A community of practice is a group of people with a shared goal, set of interests, or concerns, who come together to establish an agreed upon method or ways of working (i.e. practice) to address the goal, interests, or concerns. We discuss this definition in more detail in **Section 4**.

Box 1.1: The AI Assurance Ecosystem

The material in this box has been co-produced with the Responsible Technology Adoption Unit, part of the Department for Science, Innovation and Technology, and is based on their [Introduction to AI Assurance](#). This box provides a summary of some of the salient parts of the guidance, as understood in the context of the TEA platform, and also introduces the reader to the idea of an 'assurance ecosystem'. However, further details can be found in the full guidance.

In March 2023, the UK government set out its AI governance framework, in a report titled, *A pro-innovation approach to AI regulation* [10]. The report identifies five cross-sectoral principles to guide and inform the responsible development and use of AI, across all sectors of the economy. These principles, based on the OECD AI principles [42], are:

- › Safety, security, and robustness
- › Appropriate transparency and explainability
- › Fairness
- › Accountability and governance, and
- › Contestability and redress

The cross-sectoral principles prescribe “what” goals or outcomes AI systems should achieve, regardless of the sector in which they are deployed. The report then sets out the role of “tools for trustworthy AI”, in supporting industry and regulators to better understand “how” to operationalise these principles in practice by providing agreed upon processes, metrics, and frameworks to achieve these goals. “Tools for trustworthy AI” refers to both assurance mechanisms and global technical standards.

What are AI assurance mechanisms?

Assurance mechanisms are not unique to AI – assurance is already practised in several domains, including safety critical industries, financial auditing, and cybersecurity. Assurance builds confidence in a system, product, or process by measuring, evaluating, and communicating something about an AI system. In the context of AI, assurance can be used to measure, evaluate, and communicate whether an AI system is trustworthy, and aligned and/or compliant with the proposed regulatory principles.

There is a spectrum of AI assurance mechanisms that can, and should, be used in combination with one another across the AI project lifecycle. The list below details a sample of some key assurance techniques (complementary to argument-based assurance and the TEA framework) that organisations should consider as part of the development and/or deployment of AI systems:

- › **Risk Assessment:** Used to consider and identify a range of potential risks that might arise from the development and/or deployment of an AI product/system.

- › **(Algorithmic) Impact Assessment:** Used to anticipate the wider effects of a system/product on the environment, equality, human rights, data protection, or other outcomes.
- › **Bias Audit:** Focused on assessing the inputs and outputs of algorithmic systems to determine whether there is unfair bias in the outcome of a decision, or classification made by the system, or input data used in the system.
- › **Compliance Audit:** Involves reviewing adherence to internal policies, external regulations and, where relevant, legal requirements. Regulatory inspection is a type of compliance audit.
- › **Conformity Assessment:** The process of conformity assessment demonstrates whether a product or system meets relevant requirements, prior to being placed on the market.
- › **Performance Testing:** Used to assess the performance of a system under varying conditions. Often used as part of conformity assessments.
- › **Formal Verification:** Formal verification establishes whether a system satisfies specific requirements, often using formal mathematical methods and proofs.

Many of these assurance mechanisms are underpinned by consensus-based global technical standards.

'Standards' can be described as rules, norms, or guidelines. They are crafted to establish a dependable foundation for cultivating collective expectations concerning a product, process, service, or system as part of governance and assurance frameworks. We see examples of their implementation in industry, academia, professions, product development and service delivery.

Standards are commonly developed in a variety of ways through a *consensus-building* processes, that may be led by academic institutions, international bodies, professional associations, industry, or formally recognised Standards Development Organisations (SDOs). Standards developed by SDOs are often referred to as 'technical standards', which are developed through stakeholder-driven processes, guided by principles such as relevance, transparency, and consensus.

By establishing standardised terminology, processes, and benchmarks for the quality of products, services, or processes, technical standards help organisations maintain high standards of excellence. Consumers, in turn, gain confidence in products that adhere to recognised standards, knowing that they meet specified criteria for safety, reliability, and performance. Additionally, standards contribute to cost savings and efficiency by

streamlining processes, reducing errors, and optimising resource utilisation.⁵

All these types of standards can help to underpin and support assurance techniques and enable assurance users to trust the evidence and conclusions presented by assurance providers. Without standards we have advice, not assurance.

Why is AI assurance important?

AI assurance has three main objectives:

1. Build justified trust in AI systems.

AI assurance mechanisms can help to build justified trust in AI systems, and overcome two key organisational challenges:

- › An information problem: Organisations need to reliably and consistently evaluate whether people should trust the system.
- › A communication problem: Organisations need to communicate their evidence to other assurance users and translate this evidence at the appropriate level of complexity, so that they can direct their trust or distrust accordingly.

The value of AI assurance is in overcoming both problems to build justified trust and drive the adoption of trustworthy AI systems across the economy (also see **Next Steps: Making Fair Assurance Cases FAIRer, in Section 4**).

2. Demonstrate alignment/compliance with relevant regulation, including the UK's proposed regulatory principles.

AI Assurance requires measuring and evaluating a variety of information to show that the AI system being assured is reliable and trustworthy. This includes how these systems perform, how they are governed and managed, whether they are conforming with a technical standard or compliant with regulations, and whether they will reliably operate as intended. Assurance processes can, therefore, provide the evidence base required to prove that a system is and compliant with relevant regulations.

In the UK context, assurance mechanisms can help regulators to evaluate AI systems, and ensure that they are aligned and compliant with the proposed regulatory principles. This will help to ensure the deployment of systems that are safe, secure, and robust;

⁵ We discuss the role of standards in the TEA framework in further detail in the following guidance document (co-authored with the AI Standards Hub): <https://alan-turing-institute.github.io/AssurancePlatform/introductory-resources/standards>

appropriately transparent and explainable; fair; and to ensure that organisations have appropriate measures in place to ensure sufficient accountability and governance; and to enable contestability and redress for affected parties.

3. Enable International Interoperability.

The growth of the digital economy has increased opportunities for international trade and partnership. Major economies like the USA, Japan, Canada, Australia, and Singapore are likely to share a context-based approach to AI governance, and many companies will also need to be compliant with the EU's AI Act to enable them to operate in the EU. Embedding effective assurance mechanisms will help organisations to demonstrate compliance with relevant regulations and standards and allow for greater jurisdictional interoperability by providing agreed upon standards and metrics against which to measure compliance.


Key actors in the AI assurance ecosystem: Roles and responsibilities

There are a broad range of actors that are needed to check that AI systems are trustworthy and compliant with relevant regulations and standards, and to communicate evidence of this to others. These actors can each play several interdependent roles within an assurance ecosystem. The table below provides examples of key stakeholders and their role within the AI assurance ecosystem:

- › **AI assurance service providers:** collect evidence, assess, and evaluate AI systems and their use. Agreed standards are required to enable independent assurance providers to communicate evidence in ways that are understood, agreed on and trustworthy by their customers. Assurance may be performed by an external, independent third-party, or by an internal assurance team.
- › **Regulators:** set regulation and best practice in their relevant domains and (where required) encourage, test, and verify that AI systems are compliant with their regulations.
- › **Standards bodies:** convene actors including industry and academia to develop commonly accepted standards that can be evaluated against.
- › **Accreditation bodies:** attest to the ongoing competence, and impartiality of AI assurance services provided by third party assurance providers against international standards. This will build trust in auditors, assessors, and suppliers throughout the AI assurance ecosystem.

- › **Government:** drive the development of an AI assurance ecosystem that supports compliance with laws and regulations, in a way that does not hinder economic growth.
- › **Research bodies:** contribute to research on potential risks of AI systems, or develop new methods and metrics for assurance mechanisms.
- › **Civil society organisations:** support multi-stakeholder feedback and scrutiny on AI systems, through oversight and stakeholder convening. CSOs can also keep the public/industry informed of emerging risks and trends through external advocacy.
- › **Professional bodies:** define, support, and improve the professionalisation of assurance standards and to promote information sharing, training, and good practice for professionals, which can be important both for developers and assurance service providers.

argument, based on methods from informal logic and argumentation theory [26]. We can define argument-based assurance as follows:



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. [27]

Argument-based assurance creates a structured argument grounded in supporting evidence, known as an *assurance case*. This assurance case is the primary means for documenting how a goal has been obtained across the lifecycle of a project and supported by a process-based form of governance (as discussed above).

The structure of an assurance case is important because it helps users and stakeholders evaluate the confidence they should have in the overall argument (i.e. the level of trustworthiness).

There are three *basic elements* of an assurance case in TEA:⁸

⁸ There are additional elements beyond these three, but in this section, we focus only on these basic elements for simplicity. See the following online guidance for further details: <https://alan-turing-institute.github.io/Assurance-Platform/guidance/components>.

1. a top-level **goal claim** to be established;
2. a set of supporting **property claims** about the project or system, which collectively specify and operationalise the goal; and
3. the **evidence** that supports the individual claims and grounds the overall argument.

A toy example of these three basic elements is shown in **Figure 1.6**.

These three basic elements tie the TEA platform more closely to the process of 'reflect, act, and justify' mentioned above in relation to the process-based governance framework [22]. First, anticipatory forms of *reflection* help a team determine which ethical goals ought to be prioritised and assured. Then, *actions* taken over the course of a project's lifecycle operationalise the goal and lead to the ability to make claims made about a project or a system, which in turn create the evidence (or documentation) that can be used to *justify* the overall argument. This connection helps to demonstrate the practical leverage of TEA as a governance mechanism that enables responsible action in accordance with structured consideration of a broad range of ethical risks.

The use of assurance cases have a long history in safety-critical domains [28]. And, more recently, the use of assurance cases for the purpose of demonstrating the safety of systems in healthcare has been explored and discussed [1], [2], [29], [30].

For instance, in a recent paper, Liberati et al. present an evaluation of a project that involved a multi-site case study exploring the use of safety cases in clinical pathways [1]. Some of the benefits of using safety cases in healthcare include a) enabling a more holistic and systems-level view of risk (also see **Section 2: Fairness and Health Equity: The Unvirtuous Circle**); b) providing a new way for teams to think about risk and reflect on risk management strategies; and c) creating new forms of knowledge and understanding through upskilling. However, they also recognised challenges with safety assurance, including a) *scarce skills and resources* (e.g. safety assurance is a time-consuming process), b) *varying quality and efficacy* of assurance

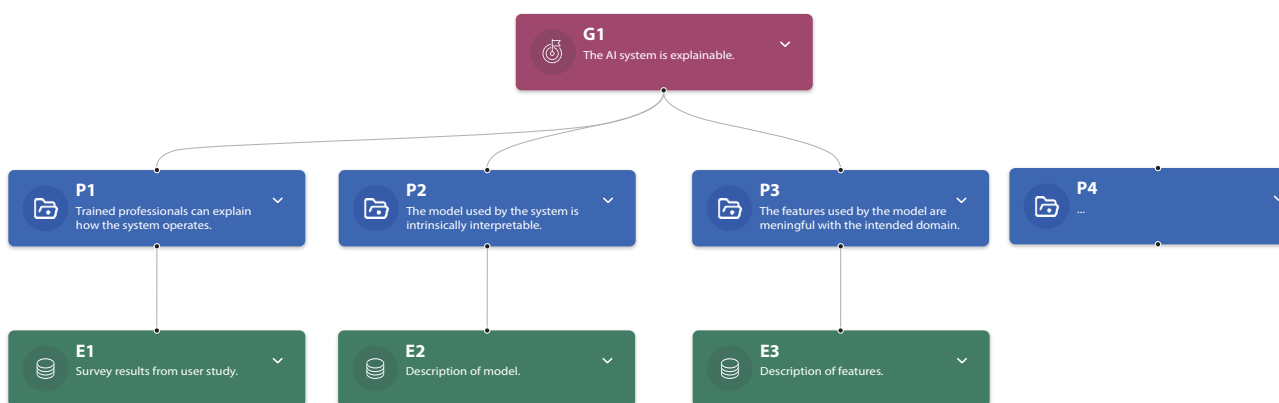


Figure 1.6: A toy example showing an assurance case for an explainable AI system.

cases can impede the reliable identification of patterns across contexts, and c) assurance cases may *uncover insights or information that may not be welcome*, especially if they lead to liability for senior leaders.

While the generalisability of their findings to *fairness* cases cannot be assumed, our own engagement activities have uncovered similar expectations to Liberati et al. (see [31] and **Appendix 1**), which suggests that the adoption of argument-based assurance as a more general methodology, inclusive of ethical goals, could help the health and healthcare sector maximise benefits and minimise risks associated with data-driven technologies. This requires us to broaden our focus beyond safety assurance.⁹

From Safety to SAFE-D: Operationalising Ethical Principles

There are many reasons why safety is of increasing interest in the context of data-driven technologies, such as AI or digital twins.¹⁰ One reason is an increased interest (or motivation) among politicians, policy-makers, and industry leaders, evidenced by the recent AI Safety Summit held in November 2023, which had as one of its objectives, the “*shared understanding of the risks posed by frontier AI*” [32].¹¹ A second (and related) reason is the growing pressure from regulators to demonstrate how safety has been established throughout a project or system, and how developers and organisations have complied with existing rules and standards [10]. And a further reason includes ongoing calls from diverse groups in society (e.g., patient representatives, healthcare professionals) to ensure that novel technologies are deployed and used safely to ensure patients and consumers are not harmed [33].

For those working in health and healthcare, the importance of safety will be well appreciated.¹² There will also be a wide appreciation of the need to approach safety as a complex and multi-faceted concept, which requires the adoption of a systems-level perspective when designing policies, interventions, and solutions [34]. As the Health Foundation notes, “safety is the product of complex interactions of attitudes, behaviours and resources.” [2] But safety is one of many goals that are important in health and healthcare.

⁹ Interested parties may also wish to read [30] for a commentary on the evaluation from Liberati et al.. This commentary includes a discussion on the changing cultural context and “patient safety mindset” within healthcare, which the authors claim is reflected in a growing interest in systems thinking, human factors and ergonomics, and resilience engineering.

¹⁰ There will be many ways to define safety, including the general understanding of safety as freedom from or absence of *unreasonable* risk, which is common in autonomous systems standards (e.g. [ISO 26262-1:2018](#)) as well as clinical safety standards (e.g. [DCB0129](#)). We don’t use any specific definition of ‘safety’ for the purpose of this report, as it is not our primary focus. However, our understanding of the term is broadly in line with such definitions, while acknowledging that safety is not a simple binary state (i.e. safe or unsafe). Our thanks to a reviewer for encouraging this clarification.

¹¹ Our focus in this report, and with the TEA platform more generally, is not limited to frontier AI.

¹² Although regretfully and too frequently overlooked in the pursuit of cost efficiencies.

Some goals and values may be in tension with safety. For instance, *cost efficiency* is obviously important in the context of a national healthcare service, such as the NHS. And even if safer technologies were (in principle) available, if the costs are too prohibitive this may end up being an over-riding factor in deciding whether to procure and deploy them. Others may be deeply intertwined with safety but emphasise a different facet of the complexity of safety. One way to illustrate this latter point is with the question, *safe for whom?*

The history of health and healthcare is replete with examples that illustrate how the benefits and risks of services, treatments, or technologies are not fairly or equitably distributed throughout society (see **Box 1.2: Examples of healthcare disparities and inequities**).

The reliability of treatments can vary for different sub-groups within a population, and safety standards for care pathways can vary within the same organisation. To illustrate this point, consider the following hypothetical scenario. An organisation developing a clinical diagnostic tool, which uses a machine learning algorithm to predict the presence of some disease, decides to carry out an evaluation of the accuracy of their system. They achieve an average accuracy of 98% across the internal and external validation of their system. They use this result as evidence in support of the claim that their system is “safe to deploy in a clinical pathway”.

Obviously, there are many issues with this hypothetical claim, including (but not limited to) the following:

- › the 2% of false predictions are all received by a specific sub-group of the population (see **Figure 1.7** for a simple illustration);
- › the representativeness of their initial dataset was insufficient to warrant generalisability of the model to new contexts; and
- › the external validation of their model granted some additional justification of the generalisability of the system, but not enough to convince stakeholders of the safety of the system overall.

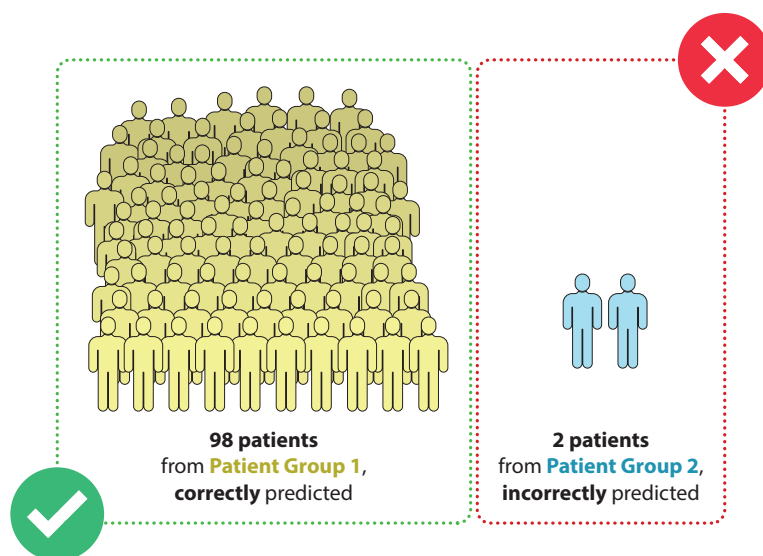


Figure 1.7: Diagram showing 100 patients, where claim about 98% accuracy is undermined by a failure to consider fairness of accuracy distribution.

Box 1.2: Examples of healthcare disparities and inequities

- **Vaccine Distribution:** During the COVID-19 pandemic, vaccine distribution highlighted significant disparities. According to a report by the World Health Organisation, “as of 1 April 2021, of the half a billion vaccines administered, 86% have been in high-income countries, while 0.1% have been in low-income countries.”
- **Genetic Research:** the majority of studies of genetic association with disease have a European bias, which has “important implications for risk prediction of diseases across global populations” and the development of genetic therapies. [35]
- **Pain Management:** there are many documented disparities in pain management, including variation in access (e.g. disparities in prescription rates) and efficacy of pain treatment [36], [37].
- **Skin Cancer:** Artificial Intelligence (AI) is increasingly used to support diagnostic processes, including dermatological applications. However, in a systematic review of open-access datasets for skin cancer images, Wen et al. found “massive under-representation of skin lesion images from darker skinned populations [38].
- **Thalidomide and Birth Defects:** Thalidomide was a widely used drug in the 1950s for treating morning sickness in pregnant women. Unfortunately, it wasn't properly tested for teratogenic effects, resulting in severe birth defects in thousands of children [39]. Despite this, the drug is still in use in some parts of the world for treatment of leprosy, and it is suggested that due to varying levels of health literacy is still causing birth defects [40].
- **Osteoporosis Treatment and Gender:** Osteoporosis is a systemic skeletal pathology characterised by loss of bone mass. Women are at greater risk than men, but evidence suggests that because of gender biases in screening, men are more likely to remain untreated and exhibit secondary osteoporosis [41].
- **Research and Evidence Gaps:** The lack of information about health outcomes can itself be a disparity. For instance, systematic gaps in understanding pertaining to the health of minority groups such as the LGBTQ+ population where some evidence suggest greater rates of mental health symptoms and disorders (e.g., depression, anxiety, substance use, and suicide) [42].

The purpose of highlighting these specific issues is to simply note where goals like safety and fairness are deeply intertwined. Therefore, demonstrating how risks have been identified, evaluated, and mitigated is, understandably, challenging. However, we argue that *Trustworthy and Ethical Assurance* provides a valuable framework and systematic approach to the operationalisation of trustworthy and ethical principles, including safety, fairness, and others such as explainability.

To help motivate this claim further, consider how the same hypothetical organisation may need to demonstrate both how their system is *safe to use* within a particular clinical pathway and also produces *explainable outputs and behaviours* that can be accessed by healthcare professionals. How they achieve this will depend on several contextual factors.

Let's take the goal of *explainability* with respect to the following questions:

- ▶ **How was the diagnostic tool developed and evaluated?** The explainability of a system is not simply a property of its outputs. In the context of healthcare and participatory decision-making, there may be a need for healthcare professionals to explain the grounds or reasons for a decision they have made, which could include reference to how a system was developed.
- ▶ **What is the level of uncertainty?** Many decisions made in healthcare will depend upon probabilistic reasoning. Presenting the results of an algorithmic decision as, say, a "90% chance of a positive result" can obscure myriad uncertainties (e.g. wide variance in a probability distribution, assumptions made during model development).
- ▶ **Who will use the tool?** Whether a tool is explainable depends, in part, on who the users are (e.g. trained professionals versus members of the public) and the design of interactions within complex systems that affect how the tool is used—a key challenge in *human factors* research.

These are just three examples of how claims made about the *explainability* of a system, in response to a small set of possible questions, are contextual and could be related to additional goals or claims (e.g. fairness and safety).

And yet, despite the contextual nature of such goals, there are also similarities in the recurring set of ethical principles that span the assurance of data-driven technologies, both within and between different domains, such as fairness and bias. In previous work [27], [31], we have explored how TEA can be used to help operationalise the following set of ethical principles:

- ➔ Sustainability
- ➔ Accountability
- ➔ Fairness
- ➔ Explainability
- ➔ Data Stewardship

We refer to these principles collectively with the acronym, SAFE-D, which is semi-recursive insofar as it is both pronounced similar to 'Safety' and also includes safety as a core attribute of Sustainability.¹³ We have also shown, in previous work, how these principles can be specified and operationalised by identifying *core attributes* that can scaffold a process of argument-based assurance by enabling teams to identify and evaluate specific sets of claims that need to be established and evidenced [45].¹⁴

Furthermore, work has also been undertaken to provide a unified argument pattern that brings complementary ethical principles together into a broader approach to assurance that focuses on the overall *ethical acceptability* of an AI system [23]. Our aim in this report is to

Box 1.3: Resources for Further Reading

Additional resources for those interested in understanding more about assurance:

- › [Safety Critical Systems Club](#): the Safety Critical Systems Club website has a wide-range of community-generated publications, including the Goal Structuring Notation (GSN) standard. GSN provides a formal definition for a graphical technique used to build and document arguments and assurance cases, and is a key influence behind the TEA platform.
- › [DSIT Portfolio of AI Assurance techniques](#): in June 2023, the Department for Science, Innovation and Technology (DSIT) launched the Portfolio of AI Assurance Techniques. The Portfolio features real-world case studies of AI assurance mechanisms being applied by organisations across a range of sectors. The Portfolio is designed to support industry, particularly start-ups and SMEs to identify relevant assurance techniques and standards for their context of use.
- › [AI Ethics and Governance in Practice Programme](#): the AI Ethics and Governance in Practice Programme comprises a series of eight workbooks and a forthcoming digital platform designed to equip the public sector with the tools, training and support it needs to apply principles of AI ethics and safety to the design, development, deployment, and maintenance of its AI systems.
- › [AI Standards Hub](#): the AI Standards Hub is a joint initiative delivered in partnership between The Alan Turing Institute (ATI), the British Standards Institution (BSI), the National Physical Laboratory (NPL), and supported by the government. The Hub's mission is to advance trustworthy and responsible AI with a focus on the role that global technical standards can play as governance tools

¹³ In other publications, we have included 'Safety' as a standalone principle, expanding the acronym to SSAFE-D (as with the example of the PBG framework above) [22].

¹⁴ As we will explore in **Section 4**, these core attributes can be used as 'strategy' elements in the process of developing an assurance case to help structure an argument.

and innovation mechanisms. The AI Standards Hub aims to help stakeholders navigate and actively participate in global AI standardisation efforts and champion global technical standards for AI. Dedicated to knowledge sharing, community and capacity building, and strategic research, the hub seeks to bring together industry, government, regulators, consumers, civil society, and academia.

- › [OECD Catalogue of Tools and Metrics for Trustworthy AI](#): there are tools and metrics out there that help AI actors to build and deploy AI systems that are trustworthy. However, these tools and metrics are often hard to find and absent from the ongoing AI policy discussions. This catalogue makes it easier to find tools and metrics by providing a one-stop-shop for helpful approaches, mechanisms, and practices for trustworthy AI.
- › [Introduction to AI Assurance](#): the Department for Science, Innovation, and Technology's *Introduction to AI Assurance* aims to build understanding of the value of AI assurance for enabling the development and deployment of safe and trustworthy systems. This guidance has been developed for a lay audience, to build baseline understanding of the role of tools for trustworthy AI, like assurance mechanisms and technical standards to support wider AI governance and operationalise the proposed regulatory principles in practice.

build on these foundations, rather than rehearsing them, with a specific focus on the principle of fairness.

A common theme across all this previous work has been the importance of recognising and understanding the societal context within which data-driven technologies are designed, developed, deployed, and used. The next section pays specific attention to this theme in the context of health and healthcare.

02



Fairness and Health Equity: The Unvirtuous Circle

Social Determinants of Health


Two Models for Supporting Fairness Assurance in Digital Health
and Healthcare

- › Looking Outwards: The Unvirtuous Circle
- › Looking Inwards: The Project Lifecycle Model

There are myriad factors that affect human health, including non-medical factors and life experiences that do not at first appear to be health related. As described below, these are sometimes described as the “social determinants” of health [46]. Acknowledging the social determinants of the health reveals that the attainment of various health outcomes is not simply a matter to be resolved by healthcare practitioners but also as a matter of social justice.

Whereas healthcare policy can improve outcomes by focusing on increasing the distributive equality of its services (e.g. ensuring that every person has similar access to the benefits of healthcare interventions), policies that promote health equity delve deeper to consider the specific requirements of individuals whose unique life paths and struggles are implicated as factors in their health outcomes.

There are many ways to define health equity. As a starting point, let’s take the following definition offered by the World Health Organisation [47]:



Health equity is the absence of unfair, avoidable, and remediable differences in health status among groups of people.

The inclusion of *avoidable* and *remediable*, here, is important. Related ethical goals such as social justice and fairness are often criticised as being lofty, poorly defined, or unachievable. Putting aside the matter of philosophical and conceptual disagreement, there is a valid point in such criticisms about how much of an ethical ideal or principle can be realistically achieved.

A person’s optimal health is the level of health they are physiologically capable of achieving given access to their preconditions of health. While it is not possible to provide every person with the same physiological conditions, it is possible (in principle) to provide a set of social conditions to support any given physiology (e.g. personalised medicine). The measure of health that results from the provision of this set of conditions, which is likely to differ from person to person, can be expressed as health equity. On this view, health equity connects the ethical values of fairness and justice to health.¹⁵

The achievement of optimal health for a society’s members is influenced by a combination of physiological and social factors. While one’s biological and constitutional features, which

¹⁵ In political philosophy, health is a feature of distributive justice, which holds that, where society can mediate the distribution of things of value, there are distributions that are more or less fair. The political philosopher John Rawls argued that an ideal society is one in which procedural fairness leads to just distributions of essential resources [90] which in turn supports flourishing. Health is presumably among the essential resources, or “primary goods” whose distribution is at least partly controlled structurally by a society and is therefore fundamental to fairness. Amartya Sen, whose own conception of human flourishing is based upon one’s capability to pursue a meaningful life, similarly argues that good health is fundamental to enacting one’s freedom, as it opens possibilities to make changes to one’s conditions of life [91]. While these perspectives suggest that health is merely instrumental to the achievement of a meaningful life – meaning it serves as a tool or pathway – health can also be considered among the implicit goals of a meaningful life; much of what we hope for is a good “quality of life” in which optimal health is a fundamental part.

are largely fixed, play a significant role in a person's ability to realise their full health potential, there are also features that are both social in origin and potentially modifiable in character. These are commonly referred to as the *social determinants of health*.

Social Determinants of Health

The WHO defines the social determinants of health as 'the societal conditions in which people are born, grow, live, work and age' [48]. The foundational theory of social determinants can be traced to public health literature of the 1970s by researchers who were critical of public health discourse limited to the study of disease progression and interventions detached from social context [49].

These conditions are shaped by the distribution of money, power, and resources (including technologies) at global, national, and local levels, which are themselves influenced by policy choices and interventions, among other things.

For instance, poverty and education are deeply interconnected with health. Lower economic status often correlates with poorer health outcomes due to factors like inadequate housing, food insecurity, and limited access to quality healthcare [50]. The famous model developed by Dahlgren and Whitehead depicts health determinants in concentric layers surrounding the individual, ranging from constitutional and personal lifestyle factors to broader social and economic conditions [51]. It is helpful for underscoring the complexity of health influences and the necessity of multi-level interventions.

For instance, wealth disparity has long been associated with suboptimal health outcomes. A systematic review by Pollack et al. [52], for example, found strong agreement among studies

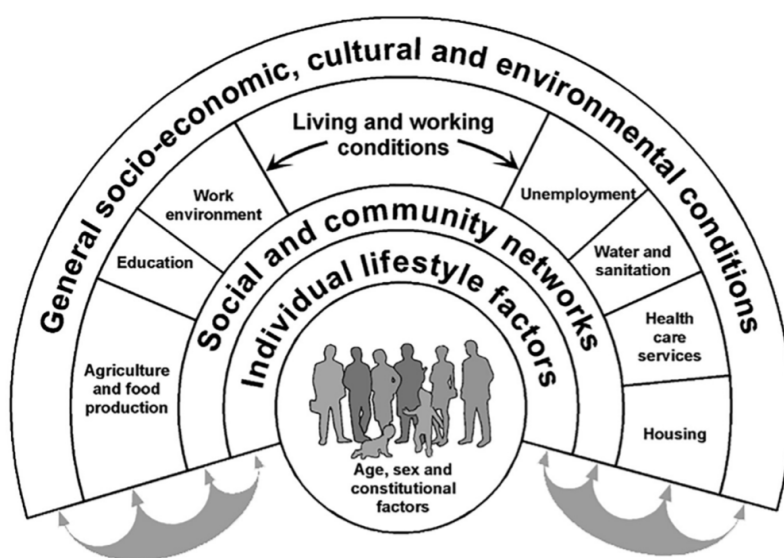


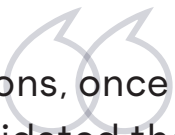
Figure 2.1: The main determinants of health (Reprinted from [51])

correlating wealth and its absence with variations in measures of health, including rates of mortality and functional status (one's ability to carry out daily or essential activities), as well as differences in self-reported health status. But while wealth may be an *indicator* of health, it is not merely the difference between rich and poor that most reliably predicts health. Daniels et al. [53] argue that degrees of health inequity are more likely to follow patterns of wealth distribution *within* rather than across societies, even among societies whose overall wealth, or lack thereof, are comparatively similar. This suggests that it is not wealth alone that determines health but various forms of social ordering. This further suggests that indicators other than income or assets are instructive to understanding the distribution of health inequity.

Several studies also demonstrate that health disparities are not limited to financial conditions. The 'Black Report' [54], for instance, provided evidence of inequity in health between racial and ethnic minorities when compared with majorities, even when accounting for economic difference. More recently, Michael Marmot and colleagues' 'Whitehall Study' documented differential health outcomes correlating the employment-grade of British civil servants; those in higher grades lived longer than those in lower grades [55]. Related to these findings, Marmot [56] identified a 'social gradient', a scale on which trends of mortality and morbidity directly correlate to a status position relative to others.

Marmot finds that health disparities are not merely a function of advantage and disadvantage. Rather, disparities in health are hierarchical throughout the gradient. Wherever analysis is conducted along the scale, health outcomes are better for those at a higher degree on the scale than those below. As Marmot found in the 'Whitehall Study', the differences were most stark between the highest and lowest grade ranks, but health disparities relative to position were evident throughout. In investigating evidence outside of the British civil service and outside of Britain, Marmot found multiple gradients with similar effects on health, including income, levels of education, class status of one's parents, and occupational prestige. While these different gradients often overlap or have tacit relationships to one another, they are sufficiently dissimilar to merit individual scales. A common thread among them Marmot identifies is autonomy; higher status as indicated by any of the gradients or some combination leads to lives with more options and freedom, which in turn improves health potential.

What has all of this got to do with AI assurance? Consider the following quote:



In many organizations, once systems engineers have verified and validated the requirements, they consider the system completed. [57, p. 92]

If fairness and health equity are inherently tied up with social determinants of health, and AI systems are sociotechnical systems (i.e. they are shaped by and in turn, shape the societal conditions in which they are situated) [58], the trustworthy and ethical assurance of such systems demands a broad (societal) perspective on the types of claims and evidence that are used to justify an argument about the overall fairness of a system.

Fairness requirements cannot be *verified* and *validated* solely from within the limited perspective of a model or system's narrowly defined technical development lifecycle.

Two Models for Supporting Fairness Assurance in Digital Health and Healthcare

All models are wrong, but some are useful.

– George Box

No single model could hope to capture all the ways in which the multitude of factors and determinants affect an individual's or population's health. Even the restricted case of how data-driven technologies interact with these factors and determinants is incredibly complex. Assumptions and abstractions are inevitable.

In this section, we present two models that we believe are *useful*, in line with the oft-quoted statement by the statistician George Box. That is, both adopt a representational perspective that constrains their utility (e.g. explanatory value or practical decision-making utility) but is nevertheless salient for their intended purpose. However, *together* they provide a more complete perspective from which to identify and evaluate how data-driven technologies are affected by underlying social inequities and injustices; inherit, propagate, and potentially exacerbate various biases; and further impact society (and its norms and practices) through their deployment and use.

Looking Outwards: The Unvirtuous Circle

Our first model for understanding data-driven technologies in health and healthcare is one that explicitly highlights the *socio-technical* nature of data-driven technologies (see **Figure 2.2**). This model offers a context that considers not only the technical features, constraints, and feasibilities of a data-driven system but also the *socio-cultural* factors that shape its design, development, and use. This four-quadrant model, known as the "Unvirtuous Circle", was originally conceived to support an investigation into questions of equity in AI-supported medical technologies [6].

The purpose of the Unvirtuous Circle model is to enable a broad perspective of the landscape of social, economic, political, and technical influences that shape the innovation ecosystem of data-driven medical technologies. This model reflects a perspective that *decentres the technology* as a singular focus for tackling the problem of fairness and equity in health and healthcare. Moreover, the model pushes against an overly narrow focus on data-driven innovations that can a) mask the importance of the social determinants of health, and b) obscure pathways to structural solutions to the longer-term patterns of structural inequality, systemic discrimination, and implicit historical bias.

Figure 2.2 provides a map with which to navigate the many entry points of inequity and unfair bias throughout the health-related AI lifecycle. Starting from the top-left quadrant and moving clockwise:

- › The **World** refers to the social and historical reality that is antecedent to the design, development, deployment, and ongoing use of data-driven health and healthcare technologies. It is the domain of lived experience in which the social determinants of health are situated and where pre-existing societal patterns of discrimination and social injustice arise. Data-driven technologies emerge from this world, and its patterns, prejudices, and attitudes can be encoded in datasets. Subsequently, they are drawn into and cascade through every stage of a project's and system's lifecycle, and risk perpetuating, reinforcing, or exacerbating existing biases and patterns of discrimination or social injustice.
- › **Data:** Attention needs to be paid to the complex social norms and practices, power relations, political, legal, and economic structures, and human intentions that condition the production, construction, and interpretation of health data. For instance, sampling biases and lack of representative datasets can be generated when individuals, who may have inequitable access to healthcare systems, distrust clinical and research environments for reason of systemic discrimination. Or, who may have limited access to digital platforms or devices (on which data are collected) and, thus, are not *accurately represented* in electronic health records (EHRs and datasets). Likewise, the source integrity of data can be prejudiced both by implicitly biased clinical judgements reflected in clinical notes, screenings, tests, and medications ordered, treatment decisions, and by inequitable tools and hardware that have been designed exclusively for dominant groups and, as a result, mismeasure minority groups. **Figure 2.2** provides an aerial view of how historical patterns of health inequity and discrimination move from the **World** to **Data**.
- › We use **Design** (bottom-right quadrant) as an abbreviation for the *sociotechnical design, development, and deployment lifecycle* of AI systems (i.e. our second model). The use of health-related AI technologies is the result of a complex and interrelated set of sociotechnical processes. As a general heuristic, these processes can be broken down into stages of (project) design, (model) development, and (system) deployment—each having a subset of activities (for instance, project design will include actions like project planning, problem formulation, and data extraction and procurement). We will say more on this quadrant when we look at the next model (**Figure 2.5**).
- › Finally, by **Ecosystem** we mean the wider social system of economic, legal, cultural, and political structures or institutions—and the policies, norms, and procedures through which these structures and institutions influence human action. Inequities and biases at the ecosystem level can steer or shape AI research and innovation agendas in ways that can generate inequitable outcomes for minoritised, marginalised, vulnerable, historically discriminated against, or disadvantaged

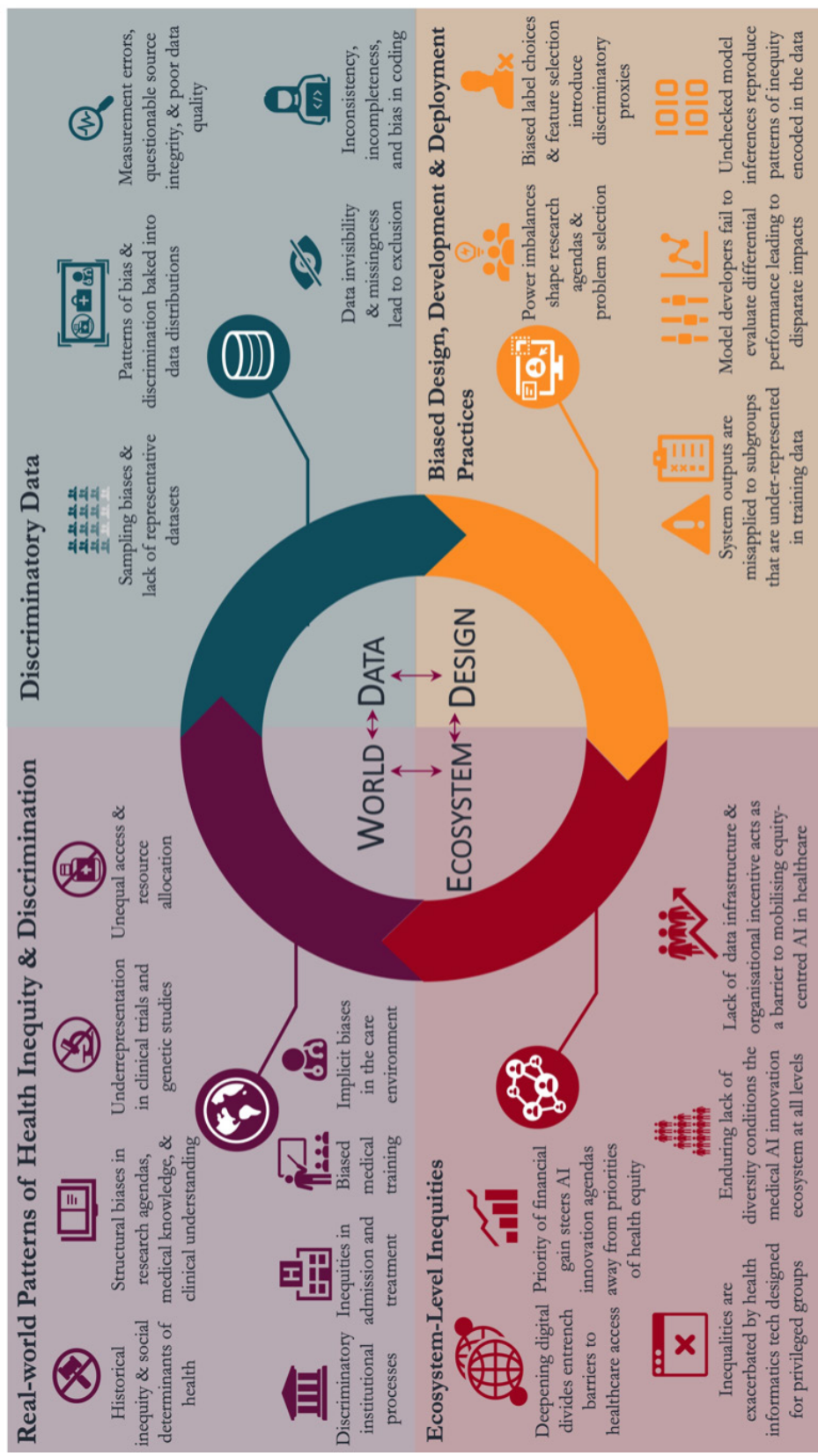


Figure 2.2: The cascading effects of inequality in the ecosystem of AI-supported medical technologies.



Figure 2.3: World to Data, movement from historical inequity to discriminatory data.

social groups. Such ecosystem-level inequities and biases may originate in and further entrench asymmetrical power structures, market dynamics, and skewed research funding schemes that favour or bring disproportionate benefit to those in the majority or those who wield disproportionate power in society. Such benefit may occur at the cost of those who are disparately impacted by the discriminatory outcomes of the design, development, and use of AI technologies.

The Unvirtuous Circle model (**Figure 2.2**) presents a *socially-situated* view of data-driven technologies. With its wide aperture, the model can help draw attention to a broad range of interlocking and interacting processes that impact upon the distribution of health outcomes in society. As such, it situates the production and use of data-driven technologies in the wider social processes that organise and shape jointly undertaken human activities, including those that lead to technological innovation.

While it can be used to support both conceptual and practical work, because of its wider perspective, it has lower specificity than a model that deals with specific stages of, say, model development. An associated challenge that emerges here is that although the model has high explanatory value—insofar as it helps identify how social biases cascade from the world, through data and design, and back to the ecosystem of data-driven technologies—it can be difficult to draw any clear boundaries regarding practical responsibility.

For instance, what should a software engineer who is firmly embedded within the development of a ML model do once they have become aware of historical inequalities in healthcare access? Is there some mitigation technique or measure they can implement to address such a concern? When should they implement such a technique? Or is the issue so removed from their work that it fails to admit a practical solution? Questions such as these are where it can be helpful to look further inwards into one of the four stages, at a lower level of abstraction.

Looking Inwards: The Project Lifecycle Model

To complement the Unvirtuous Circle model, we now turn to a second model that further splits the third quadrant of the first model into more concrete stages where practical decision-making occurs within a project's lifecycle. This is important for our focus on Trustworthy and Ethical Assurance, as it is during these stages where *claims* and *evidence* for an assurance case are established.

Figure 2.5 shows the typical stages of a project that involves the design, development, and deployment of a data-driven technology, such as a ML algorithm or an AI system. The model can be useful for several processes or outcomes, including initial *reflection* about the tasks or actions that should be undertaken at the respective stages; *anticipation* and *assessment* of significant risks, hazards, or challenges; *deliberation* about how the tasks and actions may undermine or promote relevant project goals and objectives (e.g. developing a fair classifier); and ongoing *practical decision-making* as the project unfolds and actions are documented. **Figure 2.6** shows the relationship between the two models.

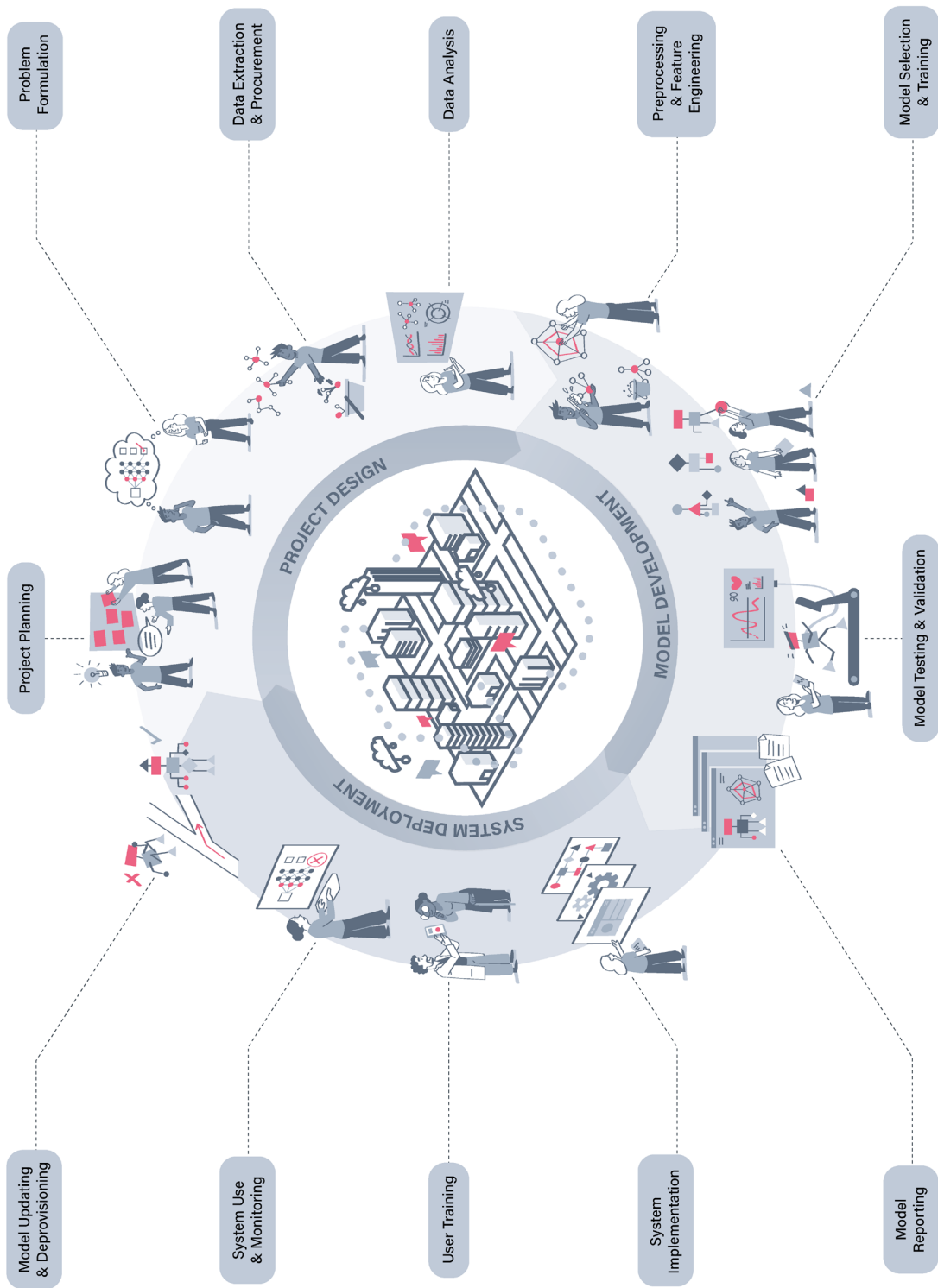


Figure 2.4: The project lifecycle model, depicting various stages across project design, model development, and system deployment.

The three over-arching stages of the project lifecycle model, in line with the third quadrant of the Unvirtuous Circle model, are as follows:

- › **Project Design:** the preliminary tasks and activities that set the foundations for the development of the model and system (e.g. impact assessments, data extraction and analysis).
- › **Model Development:** the technical and computational tasks associated with machine learning, such as training, testing, validation, and documentation, which are necessary to ensure the model is appropriate for its intended use with the target system.
- › **System Deployment:** the tasks that ensure the safe and effective deployment and use of the system (and underlying model) within the target environment by the intended users. This stage includes ongoing monitoring, as well as tasks associated with updating or deprovisioning.

There are many standards and best practices that offer similar forms of process-based governance, such as procedural standards for risk assessment [59], [60]. What is unique about the project lifecycle model, aside from its complementarity with the unvirtuous circle model, is that it has been designed to help teams embed practical forms of *ethical* and *responsible research and innovation* [61], [62] into a project’s lifecycle in the following ways:

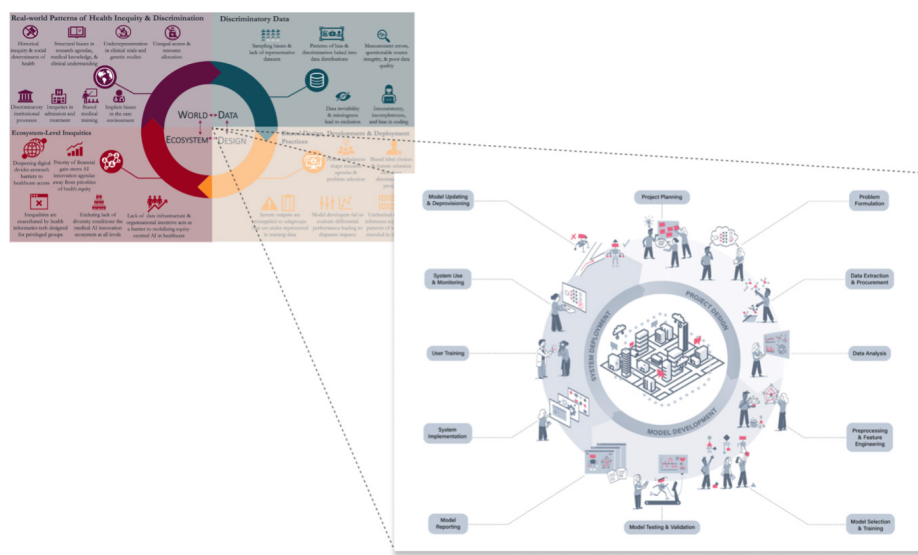


Figure 2.5: Relationship between the two lifecycle models.

-
1. **Anchoring use of RRI tools:** there have been many attempts in recent years to develop *tools and methods* for responsible research and innovation in data science and AI [20], [63], [64]. The project lifecycle model provides a unified structure for anchoring them in the practices of distributed teams,¹⁶ especially when used as part of the process-based governance framework introduced in **Section 1: Introduction** [22].

 2. **Shared vocabulary:** although a heuristic model that will not perfectly capture the nuances of project organisation across all contexts and organisations, the project lifecycle model provides a *shared vocabulary* that facilitates communication between diverse roles, stakeholders, and communities.

 3. **Framework for assurance:** when used as a framework in the context of trustworthy and ethical assurance, the project lifecycle model supports the systematic identification and documentation of claims and evidence about the decisions and actions undertaken by a team to assure a desirable goal (e.g. fairness). For instance, claims about actions undertaken during ‘model selection and training’.

We provide further details about the project lifecycle model in [27]. Here, it will suffice to highlight how the model can be used to support and structure ethical reflection and deliberation, and how it can help project teams identify or develop strategies, claims, and evidence for constructing a trustworthy and ethical assurance case.

Therefore, let us consider a hypothetical scenario.

A city hospital wants to use AI to predict and manage seasonal influenza outbreaks to ensure they have adequate staff, beds, and medical supplies. They plan to work with a team of researchers and developers to develop the system, which will use a predictive model trained on a variety of data to help inform resource allocation decisions made by the hospital staff. The system is not making *autonomous* decisions but is rather intended to *support* human decision-making, providing one source of evidence for human decision-making (i.e. a human-in-the-loop system).

Before they begin, the project team decide to carry out a preliminary risk and impact assessment, including considerations of the fairness requirements for developing and adopting such a system. How would the project lifecycle model help scaffold this process?

By systematically working through a model such as this, the team can identify specific risks that are both indexed to specific stages of a project (and its governance) and linked to actions that they could take to minimise such risks at the relevant stage. Let’s look at one example from each of the three overarching stages.

¹⁶ For instance, the bias card activity is based around the project lifecycle model. It enables teams to identify how specific biases may affect activities at specific stages of the project lifecycle, and to identify possible bias mitigation activities.

Project Design (Data Extraction and Procurement)

The team consider which data types are likely to be useful for training the model, and consider the following list:

- › Historical patient data including the number of flu cases reported each season.
- › Local demographic information.
- › Weather data.
- › Social media data and search engine trends to capture real-time public concern about flu symptoms or outbreaks.
- › Mobility data from public transport systems to understand patterns in population movement that might affect the spread of influenza.
- › Data on local vaccination rates, which can affect herd immunity and outbreak severity.

They recognise that each data type presents a possible source of bias, which could affect the accuracy of the resulting model. For instance, variation in age groups of patients using social media versus those likely to be most affected by influenza; or gaps in historical patient data such as under-representation of communities that are less likely to access healthcare. They consider various forms of data augmentation, including additional collection, to address such biases.

Model Development (Model Testing and Validation)

The team agree that a time-series forecasting model, such as a Recurrent Neural Network (RNN) or Long Short-Term Memory (LSTM) network, would be best for this problem, as such models are adept at handling sequential data and can predict future events based on historical patterns. They recognise that it will be important to not only split the data to tune hyperparameters and evaluate performance, but also to ensure that the model is validated on different time periods and, if possible, externally validated to see whether the model maintains predictive performance for a different population.

User Training (System Deployment)

Further downstream, the team recognise that a significant risk factor for the equitable impact of such a model is how well users are trained on how to use the system (i.e. *human factors*). Without sufficient training on the system, including its limitations, there is a risk of both *over-reliance* and *under-reliance* by possible users, including varying levels of uptake and usage patterns by staff. This would also be important for the subsequent stage of *System Use and Monitoring*, as members of staff may serve a crucial role in identifying data and model drift (e.g. changes in the model's performance due to a change in the relationship between the input data and the target variable, possibly due to new flu strains or changes in population

immunity). As such, they implement a suitable user training scheme or curricula to ensure the system is implemented and used correctly.

As this short example illustrates, the project lifecycle model provides scaffolding for both the reflective and anticipatory stage of project governance (i.e. risk assessment and planning).^{NEW} And, further, it can provide a shared vocabulary for structuring the ongoing and iterative governance of the project as activities and decisions unfold across the actual project.

In the following section, we explore two illustrative case studies that further demonstrate these values in the context of trustworthy and ethical assurance.

03



A Plurality of Approaches: Assuring Fairness and Health Equity

Case Study 1: AI-enabled clinical diagnostic support system

- › Context
- › Key Information
- › TEA Approach: Identifying Fairness Requirements with the Project Lifecycle Model

Case Study 2: Cardiac Electro-Mechanics Research Group Application

- › Key Information
- › TEA Approach: Developing a Fairness Assurance Case
- › Challenges and Limitations

In this section we explore two case studies, which both take fairness as a goal to be assured but approach the shared goal in different ways. Rather than undermining the coherence of the TEA platform and framework, the variety of approaches to assurance is an important feature (rather than a bug). For instance, the development and maintenance of a flourishing assurance ecosystem depends on a plurality of assurance methods, processes, actors, and organisations (see **Box 1.1: The AI Assurance Ecosystem**).

The two case studies we explore in this section are as follows:

-
1. A project involving an **AI-enabled clinical diagnostic support system** (CDSS) whose purpose is to help clinicians predict the risk of developing hypertension in patients with Type 2 diabetes.
-
2. A project involving the **design and development of a medical imaging platform** (CemrgApp) with custom image processing and computer vision toolkits for applying statistical, machine learning, and simulation approaches to cardiovascular data.

Each case study shows a different approach to TEA. The first case study uses the project life-cycle model directly, identifying “fairness requirements” that can be asked and identified at the various stages. The second case study shows the development of an assurance case that uses an operational definition of fairness to identify strategy elements that help structure an assurance case. Both case studies apply to research projects in active development. As such, it is not yet possible to provide full assurance cases for them. Instead, what is offered here are explanations of how fairness assurance cases would be developed, or are being developed, and the types of claims and evidence that would be made.

Case Study 1: AI-enabled clinical diagnostic support system

Box 3.1: Case Study 1 Project Summary

The use of AI-enabled systems in healthcare, as in all sectors, is increasing and there are continuous efforts to expand their capabilities so that they can assist or replace humans. This case study concerns the use of an AI-enabled CDSS (clinical diagnostic support system) developed by researchers at the University of York that is intended to assist clinicians in their assessment of patients diagnosed with Type 2 diabetes [65].

Context

Briefly, Type 2 diabetes occurs when a person's pancreatic cells are unable to produce enough insulin, the hormone that regulates blood sugar, or when a person's body cells do not react to insulin, leading to an imbalance in blood sugar levels. Type 2 diabetes can occur at any age and the global prevalence of this health condition has increased gradually [65]. It is predicted that over half a billion people will be diagnosed with Type 2 diabetes by 2040 [65], [66]. While Type 2 diabetes can be managed, poor management of the condition can cause the disease to progress and comorbidities, i.e., additional illnesses, to develop, such as blindness, kidney failure, hypertension and heart attacks, strokes, and the need for limb amputation, among others [65].

The early diagnosis of Type 2 diabetes is, therefore, important not only for the management of the disease itself, but also for the management of the other illnesses that diabetes can cause. This is where AI-enabled systems enter the picture. The basic idea is that AI-enabled systems can be used to advise clinicians in supporting the management of a patient's Type 2 diabetes by providing clinicians with a prediction about whether a patient is at risk of developing certain comorbidities. In particular, our focus is on an AI-enabled system that will assist clinicians by making predictions about whether a patient is at high or low risk of hypertension (e.g., within the next diabetes review) [67]. Importantly, these are patients that have already been diagnosed with Type 2 diabetes.

Key Information

To train the AI-enabled system, electronic patient data from the Connected Bradford dataset was collected. This dataset consists of real-world healthcare data from different hospitals in the Bradford, UK area. After initial filtering of the data to exclude patients without Type 2 diabetes, the data was further pre-processed such that, out of 476,000 patient records (those with Type 2 diabetes), 42,000 were selected for training and testing purposes [65].

From these 42,000 patient records, the 20 most frequent variables (e.g., body mass index measurements, total white blood cell count, red blood cell count, etc.) were included as features for the training and testing data sets [65]. That is, out of all the different pieces of information collected during a patient's visit with their clinician, e.g., blood pressure, heart rate, urea levels, etc., only 20 of the most frequently collected pieces of information were used for training and testing.

For the actual training, four different machine learning algorithms were employed: a naïve bayes, a neural network, a random forest and a support vector machine [65]. While each algorithm was used to train a separate model, the four models were combined using a generalized linear model as an ensemble method to combine the predictions of each of the four individual models and thereby increase the performance of the system [65]. The performance metrics Accuracy and Cohen's Kappa values were used as they are commonly utilised to evaluate the performance of ML-based classification models [65]. Accuracy is a metric that measures the correctness of the model, and Cohen's Kappa is a statistical measure that shows the level of agreement between the actual and predicted values by the ML model(s) [65]. Observing higher Accuracy and Cohen's Kappa values corresponds to better performance outputs [65]. In addition, to make the results more explainable, feature importance levels of each input feature have been calculated. The feature importance levels show the relative contribution of each input feature for predicting the output [65].

TEA Approach: Identifying Fairness Requirements with the Project Lifecycle Model

As mentioned in **Section 1: Introduction**, there are three basic elements of TEA case:

-
1. a **top-level claim**, e.g., that an AI-enabled CDSS used to predict whether patients with Type 2 diabetes are at risk of developing hypertension within the next six months is fair;

 2. a **set of supporting claims** that together specify and operationalise the top-level claim, e.g., that appropriate measures of statistical fairness were chosen for testing and validation of the system and that the system provides human-centric explanation of its predictions; and

 3. at bottom, the **evidence** that grounds the argument by justifying the intermediate supporting claims, e.g., documentation that indicates that the CDSS did not violate identified statistical measures of fairness during the testing and validation stage of the lifecycle.

In what follows, we outline how fairness assurance across the lifecycle would be developed, and identify questions and claims, as well as evidence, that would be made. To help illustrate

this, we have created a fairness considerations map (**Figure 3.1**) to help reason about different fairness considerations across the AI lifecycle.

Importantly, this map is not exhaustive but rather a proof of concept and guide intended to prompt further reflection and development of a fairness assurance case. Using the questions highlighted in **Figure 3.1**, we turn now to outline how fairness assurance would proceed, beginning with the design phase.

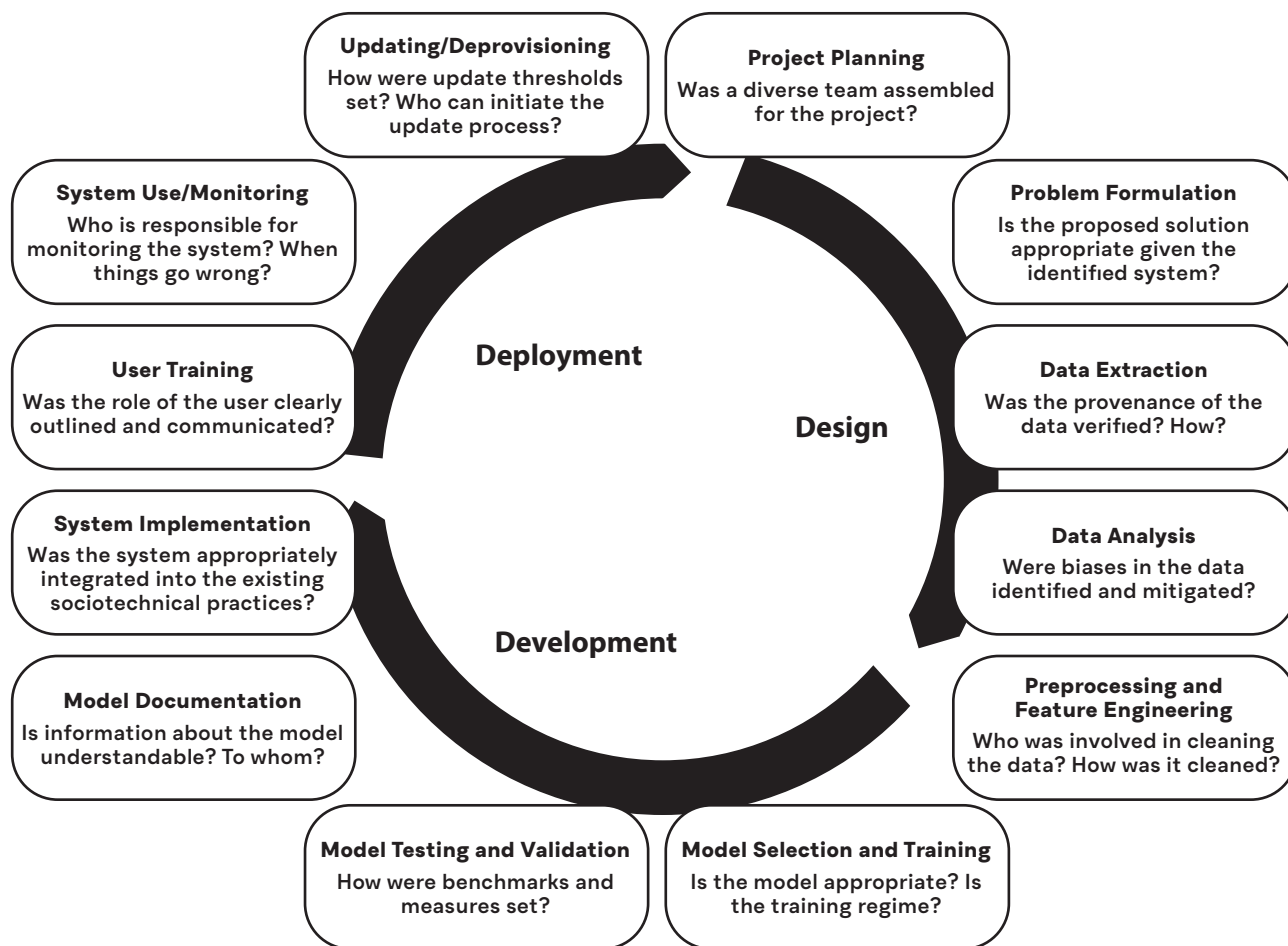


Figure 3.1: Important questions that should be included as claims supported by appropriate evidence in a TEA case intended to justify the fairness of an AI-enabled system.

Project Design

Assuring the fairness of an AI-enabled CDSS ought to begin long before data is collected, analysed and an ML model trained. It should begin in the **project design** phase, receiving proper attention in the project planning and problem formulation stages. Both stages require *appropriately diverse teams*, and evidence thereof, to justify the top-level goal that the AI-enabled CDSS is fair. Not only is a diverse project team, or inclusive forms of stakeholder engagement and meaningful participatory design—an important precondition for assuring

the fairness of an AI-enabled system—it is often required in the healthcare sector where patient well-being is held as paramount.

Beyond patient well-being, appropriately diverse project teams may also identify that an AI-enabled CDSS may have consequences for clinician well-being. The proposed solution (i.e. introducing an AI-enabled CDSS to supply an additional recommendation and thereby increase clinician accuracy when diagnosing hypertension in Type 2 diabetes patients) may introduce auxiliary problems (e.g., a distrust in clinicians or a deskilling of clinicians overly reliant on the system). Relatedly, the proposed solution may also exacerbate and further entrench harmful practices (e.g., the reduction of complex biological facts to a single measure of BMI). Furthermore, though the elimination of injustice is not guaranteed, appropriately diverse project teams can help to ensure that marginalized voices are heard and thereby alleviate some of the harms that arise as a result of epistemic injustices such as undervaluing, overlooking and misrepresenting people's testimony [68], [69], [70].

Prediction-recipients (i.e., the patients) and expert users (i.e., the clinicians) also often speak a different language than the developers. By working together with developers in the earliest stages of the project, patients and clinicians can help mitigate the perpetuation of distorted narratives about their lived realities that are often embedded in technological solutions.

It is equally important in the design phase to ensure that the data is of sufficiently high quality and that biases have been identified and mitigated in the data extraction/procurement and data analysis stages respectively. In our case, the data used was from the Connected Bradford Primary Care dataset [65]. While the dataset contains over one million patient entries, initial filtering was for those patients with Type 2 diabetes resulting in the 476,000 patient records mentioned above which was ultimately whittled down to a subset of 42,000 patient records [65]. Additionally, because not all patient records were complete, i.e., containing all of the same recorded variables from clinical visits, synthetic data was generated to fill these missing values before training of the ML model [65]. Evidence that fairness has been addressed at these stages in the lifecycle is crucial because biases in the data, e.g., over representation of a particular patient group or data of poor quality, and biases from missing elements in the data, e.g., missing age groups, can have disproportionately harmful effects on different groups of patients. Failure to substantiate claims that fairness concerns have been addressed in these latter stages in the design phase can have a cascading and pernicious effect on the fairness of the system as it is further developed in the development phase [67].

Model Development

As indicated in **Figure 3.1**, there are important elements of fairness in the **model development** phase that go beyond the merely technical or statistical. For example, justifying the fairness of an AI-enabled CDSS in the preprocessing and feature engineering stage ought to involve demonstrating that the data was appropriately cleaned by an appropriate team. In

¹⁸ So, if, for example, the BMI was missing in a patient record, it would be filled in with the average calculated from BMI's recorded for other patients.

our case, a clinician was consulted at multiple points in the preprocessing and feature engineering stage to ensure that the chosen features from the patient records align with those features a clinician uses to assess a Type 2 diabetes patient's risk of developing hypertension. Additionally, missing data was imputed as the average from other patient records in which the value was present [65].¹⁸ Justification concerning the sufficiency of this particular process would need to be included in the TEA case since there are myriad ways in which developers can choose to address the problem of missing data, each with different consequences on the fairness of the model and overall AI-enabled CDSS.

The choice of model and training regime also ought to be clearly justified in the TEA case. Too often models and training regimes are chosen because they were used by researchers investigating a similar problem and not for some other reason, e.g., because of their explainability or impact on fairness for different prediction-recipient groups. Indeed, the models for our AI-enabled CDSS, described above, were chosen because they are among the most commonly used for Type 2 diabetes-related problems [65], [71]. Testing and validation benchmarks were chosen in a similar manner for our AI-enabled CDSS. That is, the default metrics that accompany the Caret (classification and regression training) package in the programming language R were used to assess the models. In particular, the measures are Accuracy, i.e., "the percentage of correctly classified instances out of all instances" and Cohen's Kappa (or simply Kappa) which is a measure of accuracy normalized to a baseline of random chance on the dataset [65]. While there is *prima facie* nothing unfair about using popular models and default testing and validation metrics, their use stands in need of justification. And it is in the TEA case that this justification ought to be made clear.

Indeed, the entire development phase ought to be comprehensively documented and summarized in the model documentation stage of the lifecycle. Here, too, fairness considerations arise. The *appropriateness* of the mode or format of documentation to affected stakeholders' needs and epistemic competencies, for example, has consequences for their understanding of the system. There are important links here between fairness and explainability for instance.

It may be difficult to justify the fairness of an AI-enabled CDSS if a clinician does not understand how or why the CDSS arrived at the prediction that it did, as a system that limits the ability for healthcare professionals to carry out their responsibilities could be seen as having a negative impact on staff well-being. Evidence that information about the model is presented in concise and easily accessible terminology can both help justify its fairness and inspire trust in the system—it could also support a TEA case focused on the goal of explainability directly.

To make our system more explainable, the importance of each of the 20 features used to train the models was calculated to show each feature's weighted importance levels when predicting the output [67]. Ideally, even decisions about chosen explainability methods should be explicitly stated and justified in the TEA case because different explainability methods, e.g., feature relevance explanations or explanation by example approaches, can mask unfair predictions an AI-enabled system might make [72].

System Deployment

While most research on AI-enabled systems ends in the development phase, such systems are increasingly developed for commercial reasons and are therefore deployed into existing activities and practices to operate alongside other technological systems and humans. That is the intention for the AI-enabled CDSS (i.e. that it will ultimately be deployed and used in clinical settings to aid clinicians in making (more) accurate diagnoses of hypertension in patients with Type 2 diabetes), but ongoing research and development is still ongoing.

As with the previous two phases, there are important fairness claims that will require justification in the deployment phase. For example, without input from clinicians in the design phase, the implementation of an AI-enabled CDSS might clash with existing sociotechnical practices with which clinicians, healthcare providers and patients are familiar. The burden may therefore fall, unfairly, on clinicians or other affected stakeholders to incorporate the AI-enabled CDSS into their existing workflow and at best this could result in unuse or misuse of the CDSS. A similar concern arises in the user training stage, to wit, poor user training or a lack thereof may place an unfair burden on clinicians, as the expert users, to teach themselves how to operate and interact with the CDSS. Clinicians poorly trained to use the CDSS may not only mistrust the system, but that mistrust could have ramifications on patient health outcomes.

Moreover, as highlighted in the fairness considerations map (**Figure 3.1**), the user training stage is where clear roles of the clinician and AI-enabled CDSS are vital, and ought to have been clearly and unambiguously outlined during project planning, even if revisions are required at the later stage. For example, is the CDSS's prediction akin to a second opinion from a junior colleague? A senior colleague? Will the clinician be punished if they disagree with the CDSS's prediction? What if the clinician disagrees with the CDSS and they are right? What if the clinician disagrees with the CDSS and they are wrong [18]? Failure to address questions such as these can result in clinicians shouldering an unfair burden of responsibility. Indeed, there are serious concerns that the human-in-the-loop working alongside an AI-enabled system in any context, be it a clinician or safety driver in an autonomous vehicle, will serve as a "moral crumple zone" [73] or "liability sink" [19] that absorbs moral responsibility and/or liability for whole system malfunctions when things go wrong.

Fairness is also of concern in the final two stages of the system deployment phase. As alluded to above, an AI-enabled CDSS will inevitably change how clinicians interact with patients and how healthcare providers manage their services and resources. It will therefore be necessary to collect evidence that the CDSS continues to meet certain fairness goals or standards *while it is in use*. This is important for offline AI-enabled systems (i.e., systems which do not change throughout their use), but especially important for ML- or AI-enabled systems, (i.e., systems which may change their behaviour while deployed).

Failure to properly monitor the CDSS while in use can result in discriminatory outcomes for different patient groups. Similarly, failure to specify whose responsibility it is to monitor the system can place unfair burdens on clinicians since they may be expected to perform this task given their proximity to the CDSS. Importantly, while there is *in principle* nothing unfair

about having clinicians monitor the performance of the CDSS, why it is fair for them to take on such a role would have to be extensively justified in the TEA case. Perhaps clinicians will be compensated in some way for taking on this task, or will be discharged from carrying out certain other duties so that they have the time to monitor the CDSS (see [23]). Whatever the justification might be, it will have to appear in the TEA case alongside evidence that clinicians can reliably fill this system monitoring role.

The concerns just raised for the system use and monitoring stage also largely apply to the final stage, system updating and deprovisioning. In short, justifying claims that the CDSS is fair requires evidence to demonstrate that an appropriate and qualified individual or team can initiate the update process. Additionally, evidence would be required to demonstrate that appropriate update and deprovisioning thresholds have been set. As in the previous stage, expecting clinicians to alert the relevant party that the system needs updating or deprovisioning may not only place an unfair burden on clinicians, but it can impact how they treat their patients. For example, clinicians might be distracted by the performance of the CDSS and divert attention away from diagnosing their patient and instead focus on analysing the performance of the CDSS. Furthermore, clinicians may lack access to the overall performance of the system and so be unable to say whether it needs updating because of a drift in performance, for example.

Case Study 2: Cardiac Electro-Mechanics Research Group Application

Box 3.2: Case Study 2 Project Summary

In this case study, we focus on a particular pipeline for the CemrgApp, referred to as the 'Scar Quantification Tool', which facilitates, visualises, and validates the steps required for the quantification of scarred tissue in an individual heart.

The Cardiac Electro-Mechanics Research Group Application (CemrgApp) is a platform with custom image processing and computer vision toolkits for applying statistical, machine learning, and simulation approaches to cardiovascular data. The application is a joint effort of researchers, engineers, and clinicians. CemrgApp has been tailored to clinical researchers' needs and technical abilities. The project aims to accelerate clinical translation and allows users to produce automated results, while reducing variations in result caused by operator input. The main goal of CemrgApp, is to enable advanced image analysis pipelines with limited user training.

CemrgApp provides a self-contained software environment, where cardiac data visualisation and workflow prototyping are presented through a user-friendly graphical interface. CemrgApp aggregates different small apps with a specific purpose, called tools.

CemrgApp is not used for diagnosis in the clinic. At the moment, it is purely a research tool.

Key Information

The Scar Quantification Tool (SQT) takes advantage of advanced, non-invasive 3D imaging technology known as late gadolinium enhanced cardiac magnetic resonance (LGE-CMR). It is the only available method for non-invasively assessing scar tissue in the heart, particularly in the upper left chamber, called left atrium. On an LGE-CMR scan, scarred tissue appears brighter than normal, healthy tissue, making it identifiable on the images. Correctly identifying scar tissues is relevant because scar plays a role in the development of abnormal heart rhythms, a common public health concern.

The primary goal of the SQT is to identify the scar tissue on the left atrium, as captured by the LGE-CMR scan. The goal is achieved by segmenting the chamber (e.g. drawing contours to identify the left atrium within the heart) to isolate it from the rest of the heart. Then, a 3D representation of the geometry of the left atrium is generated from the segmentation. Finally, a scar map is generated, where the signal intensity from the LGE scan is projected onto the 3D model, where different colours indicate the amount of scar tissue per region. The segmentation is achieved through sophisticated deep learning approaches, i.e. convolutional neural ne-

work designed to delineate and classify different atrial structures, necessary for the analysis of scar distribution. The user design presents several push buttons, sequentially numbered to present steps in a workflow, which are then visualised within the same user interface.

TEA Approach: Developing a Fairness Assurance Case

Let's begin by asking why a fairness assurance case is important for this tool?

The CemrgApp, in general, supports patient-specific recommendations. In other words, it is tailored to an individual patient based on their data. While this brings many opportunities for personalised healthcare, it also carries risks. Most notably, the possible risk of unequal performance for individuals or sub-groups of the population. Therefore, it is worth considering how such risks can be identified, evaluated, and mitigated.

We approached the development of an assurance case for the SQT using the following general approach:

-
1. explore a general understanding of the **concept of fairness** with members of the project team;
 2. consider different **examples of practical fairness issues** that are present (e.g. representativeness of dataset, biases in design and development); and
 3. use **core attributes of fairness as strategies** for identifying exemplary claims and evidence for a draft assurance case.
-

This approach diverges from a more comprehensive approach that would be required if a full assurance case were being developed and published, but as you can see below, this approach still provides a useful framework for reflection, deliberation, and the initial development of a fairness assurance case.

Starting with (1), there was an initial challenge around the scope of the assurance case. For instance, we considered the following question:

Which component are we focusing on? Or, how are we delineating the boundaries of the system?

CemrgApp is multi-functional and has been designed to support multiple clinical use cases. And so, assuring the whole app was too challenging for this initial case study.

Therefore, it was decided to focus on a specific tool and use case—the SQT for diagnosis and clinical decision making— to help focus reflection and deliberation. This narrowing of context

is fairly common practice in assurance (i.e. setting constraints or parameters on the validity of the assurance case), and assurance elements can help clarify where such constraints exist (i.e. context claims).

Turning to (2), we considered which core attributes of fairness and health equity should be emphasised. Given the focus of this project, this higher-level goal was fixed in advance.

The team settled on the following four core attributes:

- › Bias Mitigation (across project lifecycle)
- › Diversity and Inclusivity (for project governance)
- › Non-discrimination (in model outcomes)
- › Equitable Impact (of system)

These core attributes are based on extensive desk research and had been refined and co-developed in previous projects that included stakeholder engagement and participatory design of the core attributes [31]. As such, they serve as a useful starting point for operationalising fairness in a health and healthcare context, even if additional claims or evidence may need to be added to ensure the completeness of an argument. For this project, they were used as *strategy elements* to help identify relevant claims and associated evidence for the development of the assurance case.¹⁹

In stage (3) the following claims were identified as initial examples for the four strategies/core attributes. For each claim, additional information is also provided for explanatory purposes, including a discussion of the evidence that was considered or identified.

Bias Mitigation

➔ Claim 1

The SQT reduces undesirable operator variability to limit the impact of cognitive biases when using the image processing pipeline and chosen thresholds.

In the current scientific and clinical landscape, the analysis of LGE-CMR imaging for scar identification is challenged by the absence of standardized analysis protocols. This situation leads to inconsistent interpretations of scar among clinicians, because when faced with variable image acquisition and processing methods, they might base their decisions on different cognitive biases and heuristics. In general, therefore, unnecessary degrees of freedom offered by a software tool (or poorly considered UI) can potentially exacerbate existing cognitive biases.

¹⁹ It is important to draw attention to the difference between this approach and the approach taken in **Case Study 1**, where the project lifecycle model was used to structure the identification of claims. It remains an open question whether these two approaches can be fully integrated or whether one or the other is preferable for a given project.

However, minimizing cognitive biases, and as a result undesired variability in clinical decisions, needs to be balanced with the accuracy of the system. A highly standardized semi-automatic processing tool is only valuable if it provides more accurate diagnosis compared to manual processing.

The SQT consists of a semi-automatic processing pipeline, which relies on specific, standardized analysis methods and objective thresholds. This standardization is instrumental to ensure clinical decisions are driven more by objective data than individual interpretations.

A reproducibility study provides empirical evidence for successful standardization of the pipeline [74]. The study resulted in consistent reproducibility, both when different observers use the tool (interobserver reproducibility) and when the same observer uses it on different occasions (intra-observer reproducibility). Reproducibility is a critical factor in reducing cognitive biases. Furthermore, the tool's reproducibility was found to be superior to traditional visual assessment, meaning that existing biases are actively minimized rather than simply maintained.

To ensure the demonstrated reproducibility levels, we assume any operator to go through an initial training phase. For this the developers supply an accompanying instruction manual for running the tool. Training is essential for effective and standardized use of the app.

As the software evolves with new features or parameters, its impact on variability and bias mitigation might need reassessment. New complexities could alter the balance between standardization and user flexibility, affecting the app's reproducibility and bias mitigation effectiveness.

Diversity and Inclusivity

Two related (but separable) claims were identified for this core attribute:

➔ Claim 2

The tool supports patient engagement and participatory decision-making through accessible and informative visualisations.

➔ Claim 3

The user interface and dashboard are intuitive and follow best practices for presentation of information.

Starting with claim 2, the visualisations afforded by the SQT are both informative to healthcare professionals but also engaging for patients. As such, use of the tool promotes par-

ticipatory decision-making between, say, a clinician and a patient—promoting inclusivity in healthcare.

At present, this claim does not have any evidence to justify its validity. However, we identified that when it comes to participatory decision making, surveys and focus groups could be used to gather qualitative data on patients' understanding and perceived empowerment from seeing and discussing the tool's visualisations with their clinician. Additionally, A/B testing could compare different versions of the visualization to determine which features or designs best facilitate patient engagement. Quantitative metrics, such as the time taken to decide or the number of questions asked during consultations, can also offer insights into the tool's effectiveness in enhancing patient participation.

Whereas claim 2 focuses on inclusivity of patients, claim 3 addresses a similar need to consider diversity and inclusivity among healthcare professionals (i.e. diverse user base for the tool).

A standard operating protocol (SOP) was co-created together with clinicians to ensure the tool is usable. The default colour schemes are chosen in line with research community standards for intuitive interpretation of the visualisations by clinicians.

These activities and design choices are necessary but not sufficient elements for evidencing the claim that the tool's user interface and dashboard is intuitive and enables participatory decision making. The team has identified a need for collecting direct feedback from a representative sample of end-users to report on the usability of the tool and the value of the SOP. This should best be done through formal usability testing, given that voluntary feedback reporting through current GitHub community features (e.g. issues and pull requests) has posed a barrier to user engagement in the past (see **Challenges and Limitations** below).

Equitable Impact

Again, two related (but separable claims) were identified here—both of which are framed in terms of the SQT but, more generally, apply to CemrgApp as a whole:

➡ **Claim 4**

The tool is open-source and easily accessible to allow clinicians to run the software.

➡ **Claim 5**

The tool is efficient in terms of computational resource, allowing for widespread use (e.g. not dependant on expensive and proprietary cloud infrastructure).

The relationship of these claims to the core attribute of 'equitable impact' may not seem immediately clear. However, it is important to note that access to software is not also equitably distributed across healthcare service providers. As such, these properties of the tool (and CemrgApp) seem worth making.

Regarding the former claim, CemrgApp's status as an open-source tool is clearly demonstrated by its availability on GitHub—a widely recognized platform for hosting and sharing open-source software. The presence of the full app on this platform not only confirms its open-source nature but also ensures its accessibility to clinicians and researchers worldwide.

The repository is published under a 3-clause BSD license, which legally permits users to freely use, modify, and distribute the software. This license is a key component of open-source software, ensuring that the app remains accessible and modifiable by the community.

Currently, the team are considering which metrics would be suitable to demonstrate the assumed validity of the second claim.

Non-Discrimination

➔ Claim 6

The tool does not discriminate across different patient groups and ensures quality and consistency of results across users with different levels of training.

The tool makes use of a deep learning technique, specifically multilabel convolutional neural networks, to automatically generate an anatomical segmentation based on the LGE-CMR 3D images [75]. Deep learning techniques are at risk for picking up on statistical biases in the input dataset and potentially propagating these biases in their outputs.

The team has taken some measure to limit statistical biases in the input data by balancing patient demographics in the input data on sex (female 42%). This approach is particularly crucial given the recognized shortfall in data on female cardiac data, mitigating an existing bias in the clinical literature. In the future, accuracy of outputs should be validated for more protected characteristics to ensure non-discrimination across other potentially vulnerable patient groups such as elderly populations, and those with certain comorbid conditions.

One potential challenge is the continued assurance of this claim. To further improve the accuracy of the neural network, the developers are continuously fine-tuning the model weights with batches of manually labelled data, as it is made available through partnering clinicians. Such an iterative development requires a robust framework for continuous evaluation to ensure that updates do not introduce new biases. The lack of diverse validation datasets currently presents a significant hurdle in benchmarking the tool's non-discriminatory performance across different patient demographics.

Example Assurance Case

With these claims and evidence in mind, we can start to put a fairness assurance case together to show the emerging argument in a graphical manner (see **Figure 3.2**).

This assurance case is clearly incomplete, both in terms of the identified claims and the supporting evidence. However, as noted, the current structure was both helpful in scaffolding the process of reflection and deliberation and in providing a means of capacity building for a team that were previously unfamiliar with argument-based assurance.

Challenges and Limitations

CemrgApp is an ever-evolving platform with continuous technical enhancements and additional tools and pipelines. Nevertheless, we acknowledge various challenges and limitations in terms of fairness assurance.

Firstly, we recognise significant gaps in our ability to substantiate claims, particularly in the systematization of feedback from end-users. CemrgApp is hosted on GitHub, providing free access to features like issue logging and discussion forums for user engagement. Despite efforts to categorise and prioritise user-logged issues, the adoption of these tools by end-users for effective communication and evidence of feedback remains a challenge. A potential reason for this is that some users may find it more convenient to convey their thoughts through personalized communication channels, such as emails or direct messages, rather than creating and managing a GitHub account for engagement with the platform.

Secondly, validating equal accuracy rates across demographics poses a systemic challenge for fairness assurance. We presented the Non-Discrimination claim for automatic segmentation in the Scar Quantification Tools, maintaining a 42/58% split for sex in the input data. However, validating accuracy rates across other demographics is challenging. Organisations aiming to ensure non-discrimination may struggle to obtain sufficient data due to cost implications and potential negative impacts of data sharing agreements between research organisations.

CemrgApp strives for continuous improvement, however, challenges in user engagement and demographic validation underscore the need for adaptable strategies and enhanced communication channels for comprehensive fairness assurance.

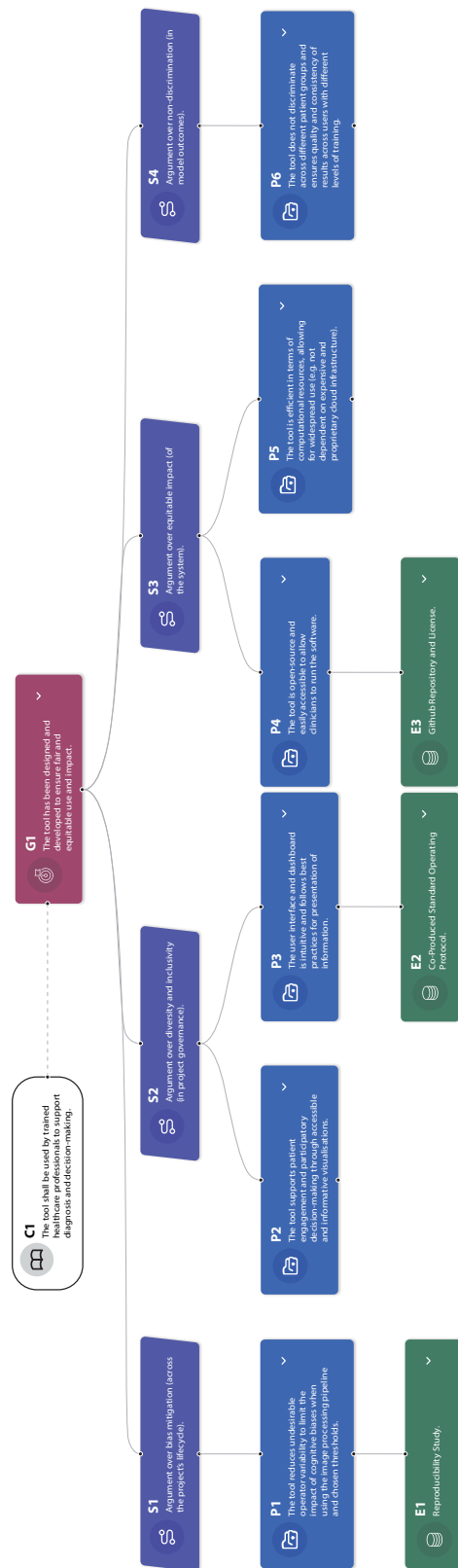


Figure 3.2: A partial assurance case for the CemrgApp's Scar Quantification Tool.

04



Scaffolding Communities of Practice

What is a Community of Practice?

Addressing Bias in Assurance Cases: The Case for Public Reason

Next Steps: Making Fair Assurance Cases FAIRer

- › Why do the FAIR principles matter in the context of the TEA Platform?

In this final section, we explore why it is important that the TEA platform is developed and maintained through an open community of practice, and what can be done to facilitate this goal. Let's first look at what we mean by the term 'community of practice'.

What is a Community of Practice?

We can think of a community of practice (CoP) as a group of people with a shared goal, set of interests, or concerns, who come together²⁰ to establish an agreed upon method or ways of working (i.e. practice) to address the goal, interests, or concerns. As noted by Li et al. [76], building on the original work of Lave and Wenger, CoPs provide a critical environment for knowledge exchange, creative collaboration, and informal interactions that build trust and rapport among participants.

CoPs can be thought of as *informal learning organisations* (including those that emerge spontaneously or are deliberately developed and managed) [76]. This is important in the context of TEA for digital health and healthcare. As Sujana and Habli [30, p. 1] note, when reflecting on the state of patient safety research back in 2008, "[m]uch effort was dedicated to the implementation of incident reporting systems and counting (rather than learning from) the number of incidents and adverse events." While significant changes have occurred in the intervening years, there is still plenty of room to improve and enhance existing capabilities, especially given the disruptive effects of data-driven technologies on healthcare, such as AI and foundation models [30], [77].

Given the inherent variety of CoPs (both in terms of their function and composition), Li et al. acknowledge that the structure of CoPs can be inconsistent, ranging from informal networks, support groups, or even just a multidisciplinary team [76]. However, as the value of CoPs become increasingly recognised by practitioners across the domains of research and development, more approaches to forming and managing CoPs are being widely shared and collaboratively developed.

For the TEA platform and framework, the purpose for considering this topic is to explore how a CoP (or diverse and networked CoPs) would enhance and facilitate our approach. Our rationale for doing this is to (partially) address some of the challenges that emerged during a series of stakeholder engagement events that were conducted as part of the TEA-DH project.

The cross-cutting themes that were identified during this engagement were as follows (see **Appendix 1** for full details):

²⁰ Here, "come together" can be understood as inclusive of options such as in-person/remote/hybrid, synchronously/asynchronously, and geographically centred/geographically distributed.

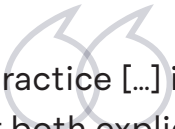
-
1. **Awareness of context** is vital when carrying out assurance activities, especially when considering goals such as fairness. Here, context may refer to the broader regulatory context, the research and development context, the use context, or the ethical context, among other things.

 2. There is a pressing need to **cultivate a flourishing assurance ecosystem**, in which many different techniques, tools, processes and standards are available (see **Section 2: Fairness and Health Equity: The Unvirtuous Circle**). However, a shared vocabulary—especially around common goals, values, or objectives—should underpin this plurality.

 3. There are many challenges posed by **current gaps in skills and capabilities**, which will create barriers to the adoption of responsible research and innovation practices. Organisations should continue to invest resources and build capabilities through upskilling activities and strategic planning.

Each of these themes connects to a challenge that can be (partially) overcome or addressed through the development of an open community of practice that is centred upon trustworthy and ethical assurance. Let's see how.

Starting with theme (1), consider the following quotation:



integrating evidence into practice [...] involves a complex process of acquiring and converting both explicit and tacit knowledge into clinical activities [...] However, to apply this knowledge in practice, practitioners must **make sense of the concrete information in the context in which it is used**. This process of establishing meaning can be facilitated by discussions with colleagues and mentors or by observing how others apply the knowledge and then try it themselves. [76, pp. 2, emphasis added]

Here, we see a direct recognition of the importance of situated learning within a community of practice as a means for supporting the application of “concrete information” to the practical actions and decisions that constitute the design, development, and deployment of data-driven technologies. Such concrete information may be codified in standards, argument patterns, frameworks of principles, and other guidance or best practice documents.

Next, theme (2) is a sensible recognition of the importance of diversity and plurality—a recognition that is directly connected to the importance of context as a means for supporting the identification of (contextually-relevant) tools and techniques.

Here, it is important to recognise that while the term, ‘community of practice’ may suggest something that is singular or homogenous, it would be false to assume that this is what is intended. Instead, we could refer to networked ‘communities of practice’, to recognise the interconnected nature of a lot of knowledge exchange and learning that happens within such communities. This would also help to emphasise the value of *open knowledge exchange* between communities as a potential means for a) critically evaluating the strengths and limitations of various tools or techniques that may be dominant in one community, b) reducing the negative impact of social and cognitive biases (e.g. group think), and c) building a more resilient and robust ecosystem—in line with the proposal laid out in **Section 2: Fairness and Health Equity: The Unvirtuous Circle**.

Finally, turning to (3) there are many challenges posed by current gaps in skills and capabilities pertaining to assurance of digital health and healthcare technologies, which will create barriers to the adoption of responsible research and innovation practices. Organisations should continue to invest resources and build capabilities through upskilling activities, including the promotion and support of communities of practice. In line with the idea that a CoP is a “learning organisation”, it is important to consider how an assurance CoP will support capability-building.

There are several concrete ways in which the TEA platform and framework have been designed to support this:

1. TEA cases are **shareable**: this is important, because a) assurance in general needs to be communicated with relevant stakeholders or end users, and b) open sharing of assurance cases allows others to learn from the work of others and to collaboratively develop best practices within a respective CoP.

2. The TEA platform is **open-source**: the underlying tool has been developed and made accessible as a public and open-source tool (available through the project’s GitHub repository). This allows the community to a) contribute to the development and maintenance of the tool (e.g. creating issues or opening pull requests), and b) fork the original code if they wish to extend its functionality within a specific context.

3. The TEA framework is **flexible**: while we have our own set of principles that can be used as top-level goals, the actual TEA framework is agnostic to the choice of normative framework. That is, a project team can use the tool and framework with their own principles (e.g. organisational goals), and also adapt the tool to use within the context of a particular form of project governance.

4. The TEA platform promotes **collaboration**: by developing the TEA platform around existing collaborative infrastructure (i.e. GitHub), the TEA platform enables teams and individuals to collaborate on the development and evaluation of assurance cases using open research and innovation infrastructure that will already be familiar to many. For instance, teams can a) import assurance cases from existing GitHub repositories, b) track changes to their assurance case using GitHub's features (e.g. version control), and c) share and collaborate with others (e.g. managing open communication and review through issues).

While the value of each of these four properties is, hopefully, clear. There is a further (and more philosophical) set of properties that require additional attention.



Further information


The above four points are documented in more-depth on our project's repository and project pages:

- › GitHub repository: <https://github.com/alan-turing-institute/AssurancePlatform>

Addressing Bias in Assurance Cases: The Case for Public Reason

In this section, we discuss the potential for *open communities of practice* to mitigate the negative effects of social and cognitive biases (e.g. confirmation bias, availability bias, selection bias).²¹


In [79], the Assurance Case Working Group (ACWG) acknowledge an important limitation of assurance cases:



Assurance cases are occasionally criticised for being biased in the way that they present their argument and evidence, and this criticism may very well be justified. Assurance cases do not, typically, present balanced arguments in the same way as one would expect to see in a hearing in a court of Law. That is, assurance cases rarely, overtly, contain an argument ‘for’ and ‘against’ the claim being made (e.g. they would typically instead argue only that ‘System A is safe’, or ‘System B is secure’).

This is important. As the above quote acknowledges, if an assurance case is akin to an argument presented in a court of Law, it is only *one of the sides* being presented. Later in their guidance, the ACWG consider how to address this possible bias (e.g. encouraging a hazard or threat-seeking culture, seeking disagreement, discussing counter-evidence) and put forward a dialogical form of argumentation that is incorporated into their GSN standard [80]. All their proposals are sensible and help raise the prospect of presenting more impartial assurance cases. There is, however, a further extension to these considerations—the *court of public reason*.

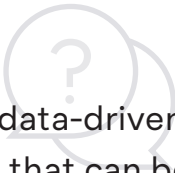
In [81], we discussed the idea that open and dialogical forms of assurance were an important means for addressing bias when responding to the concern that ethical assurance cases could be misused as forms of “ethics-washing” (i.e. where selective evidence is presented to overstate the ethical permissibility of a project). Our response was as follows:



By exposing the argumentation structure [of an assurance case] to open critique and active enquiry, an ethical assurance case is more likely to expose an unconvincing or incomplete argument. In turn, there is a potential for improving the argument, or using available legal mechanisms to hold the organisation accountable. [81, p. 22]

²¹ See [78] for a structured activity around the identification and mitigation of biases across the design, development, and deployment lifecycle of data-driven technologies, such as AI.

The extent to which a project team may be able to open an assurance case up to critical evaluation in a public setting (or *permitted to*, given other interacting duties or obligations, such as data privacy or intellectual property) will, of course, vary (e.g. risk of leaking sensitive information, enabling malicious users to exploit or game a system). However, in general, we would argue that presenting an assurance case to open critique and evaluation would be beneficial to encouraging a trustworthy and responsible approach to research and innovation. As an illustrative example to support this claim, and sticking with our over-arching theme of fairness in the context of data-driven technologies, consider the following question:



What does it mean for a data-driven technology to be fair or equitable in terms that can be operationalised?

Many statistical measures of fairness have been offered in response to this question, and while intuitively plausible a key problem is that they often cannot be simultaneously satisfied [82]. Ruf and Detniecki [83] have developed a helpful tool, known as the Fairness Compass, which shows some of the epistemic and normative assumptions that underpin the possible choices made between such statistical measures (e.g. whether there are equal base rates between the relevant groups), and other similar tools or software libraries exist for similar purposes [64], [84]. But such tools do not obviate the need for hard choices to be made and justifications about such choices to be presented.

In presenting justifications for choices or actions made throughout the process of designing, developing, and deploying a system, we are in effect giving (and defending) reasons that we expect others to accept. And, if we are unable to provide such reasons, this may be a good sign that our choices are *not defensible in the relevant setting* (e.g. public sphere). We can think of this in terms of moral and political philosophy, such as the theory of contractualism, where it is claimed that the content of such reasons ought to be based upon concepts and principles that would be acceptable to rational agents engaged in public justification.[85]

While the specific strategies and claims in most assurance cases will fall far short of such a normative ideal, we can still strive for adherence to the relevant values in the overall development and communication of assurance cases (e.g. legitimacy, impartiality, consistency, and mutual respect). For instance, consider the over-arching ALARP principle for risk management in safety cases which claims risk should be reduced “as low as reasonably practical” [86], balancing what’s reasonably achievable with the effort and resources required. While there will be plenty of room for reasonable disagreement in terms of specific claims and evidence, ALARP serves as an over-arching and governing principle in safety assurance and is something that should be accepted by all reasonable people. So too, should we aim for shared communicative values that govern the production and communication of trustworthy and ethical assurance.²²

²² We have elsewhere presented a framework for this broader form of governance, focusing on the public sector [22].

Next Steps: Making Fair Assurance Cases FAIRer

The TEA framework is an ongoing project. So, rather than a “conclusion” it seems fitting to end this report with a forward-looking discussion of next steps and open questions. One key piece relates to the consideration of how the TEA platform and assurance cases adhere to the FAIR principles (and why this matters).

Why do the FAIR principles matter in the context of the TEA Platform?

As **Box 4.1: What are the FAIR principles?** acknowledges, the FAIR principles may have emerged initially as a set of best practices for managing data and associated metadata, but they

Box 4.1: What are the FAIR principles?

The FAIR principles (Findability, Accessibility, Interoperability, and Reusability) are a set of guidelines aimed at enhancing the management and governance of data. Collectively, they emphasise the importance of making data easily and openly available for both human users and computers (e.g. ‘interoperability’ to allow software to effectively interpret, process, and integrate data from various sources). The core idea is to ensure that data are collected, processed, stored, and structured in a way that maximises their potential use and impact.

The “How to Fair” website introduces the following four statements [87], which serve as clear motivations for why the FAIR principles are important:

1. Both humans and machines are intended as digesters of data.
2. The FAIR principles apply to both *data* and *metadata*.
3. The principles are not necessarily about open data.
4. The FAIR principles are not rules or standards.

The FAIR principles have been widely cited and adopted, with many expanding the initial framework and reach of the community by developing supporting tools and resources (e.g. resources (e.g. the Data Stewardship Wizard for creating Data Management Plans that adhere to FAIR best practices [88]). Such tools can help projects at different stages of a project’s lifecycle, either by helping with the “FAIRification” of data (e.g. automating documentation or consistent metadata), or by enabling teams to answer reflective questions such as “how FAIR are we currently?” [87], [89].

are now broader than this. For instance, “data” can refer to all kinds of digital objects that are produced in research, including code, models, software, presentations, etc. This is promising, but also raises a key challenge, as articulated by Thompson et al., “Even though it is clear what is required by these principles, it is not specified how it should be done, i.e., FAIR is not, in itself, a standard.” [89, p. 88]

Because assurance cases are *structured data*, they fall under the remit of the FAIR principles. However, a stronger argument can be made for why they should be governed by them. Let’s consider this argument alongside each of the four FAIR principles to see how we can make progress with answering the above challenge referenced by Thompson et al.

1. Findability

- › **Metadata and Standardised Tagging:** Implementing detailed metadata for all data used in assurance cases can enhance findability. This includes tagging evidence, claims, and arguments with standardised and searchable tags (e.g. claim about ‘bias mitigation’, ‘evidence related to a particular standard’), making it easier to locate relevant information.
- › **Open (and Public) Repositories:** Establishing centralised repositories where assurance case data is stored, indexed, and searchable can significantly improve the ability to find necessary information.

2. Accessibility

- › **APIs for Data Access:** Developing and maintaining APIs that allow for programmatic access to assurance case data can facilitate automated systems in retrieving and processing this information.
- › **Clear Access Control Policies:** Implementing clear, well-documented access control policies ensures that data is accessible to authorized entities while maintaining security and confidentiality where required. At present, the TEA platform supports basic sharing (e.g. importing/exporting) and can be deployed within an organisation using our source code (or public Docker images). However, further work will be required to ensure access control is fit-for-purpose across different contexts (e.g. in organisations that make use of sensitive information).

3. Interoperability

- › **Standardised Data Formats:** Utilising standardised data formats (e.g., XML, JSON) for assurance case data ensures that different tools and systems can easily exchange and process this data. Currently, the TEA platform uses JSON as the primary data

format, but this diverges from other approaches, such as Structured Assurance Case Metamodel (SACM)—a framework defined by the Object Management Group (OMG) for representing structured assurance cases—that proposes the use of XML documents that conform with the SACM XML Schema.

- › **Common Frameworks and Ontologies:** related to this previous point, adopting common frameworks and ontologies for assurance cases promotes a shared understanding and compatibility between different systems and organisations. The TEA platform has not, hitherto, aimed for the same expressivity as other frameworks or ontologies (e.g. GSN and SACM), because we have prioritised accessibility of arguments for non-technical stakeholders to promote a more accessible and inclusive assurance ecosystem, which is vital to TEA. However, future research will explore how convergence with other frameworks can be achieved, and how much is desirable.

4. Reusability

- › **Modular Design of Assurance Cases:** Designing assurance cases in a modular fashion, where components like evidence, claims, and argument structures can be reused in different contexts, enhances reusability. The GSN community standard supports modular design of assurance case through their 'Modular Extension' (see section 1.4 of [80]). Some of our work has also looked into applying this modular design to ethical assurance cases [23]. While this feature is not currently available in the TEA platform, it is a key milestone and priority to improve reusability of assurance cases (e.g. as modules within larger scoped arguments).
- › **Comprehensive Documentation:** Ensuring comprehensive documentation is vital if knowledge that can be derived from assurance cases is to be reused—perhaps in different domains or contexts. Therefore, in addition to the information contained within the structured assurance cases, it will be important to consider how assurance cases can be accompanied, say, by short summaries to help stakeholders understand the case.

These comments about the FAIR principles, as they apply to the TEA platform, also help bring us full circle to a point raised in the introduction. There, we noted that Liberati et al., in their scoping research of the use of safety cases in healthcare, stated that there were *limited empirical studies* or case studies exploring successful use of safety assurance in healthcare [1]. This challenge also applies to the TEA framework. In fact, it is felt more acutely due to the novelty of the use of argument-based assurance to operationalise and justify trustworthy and ethical goals.

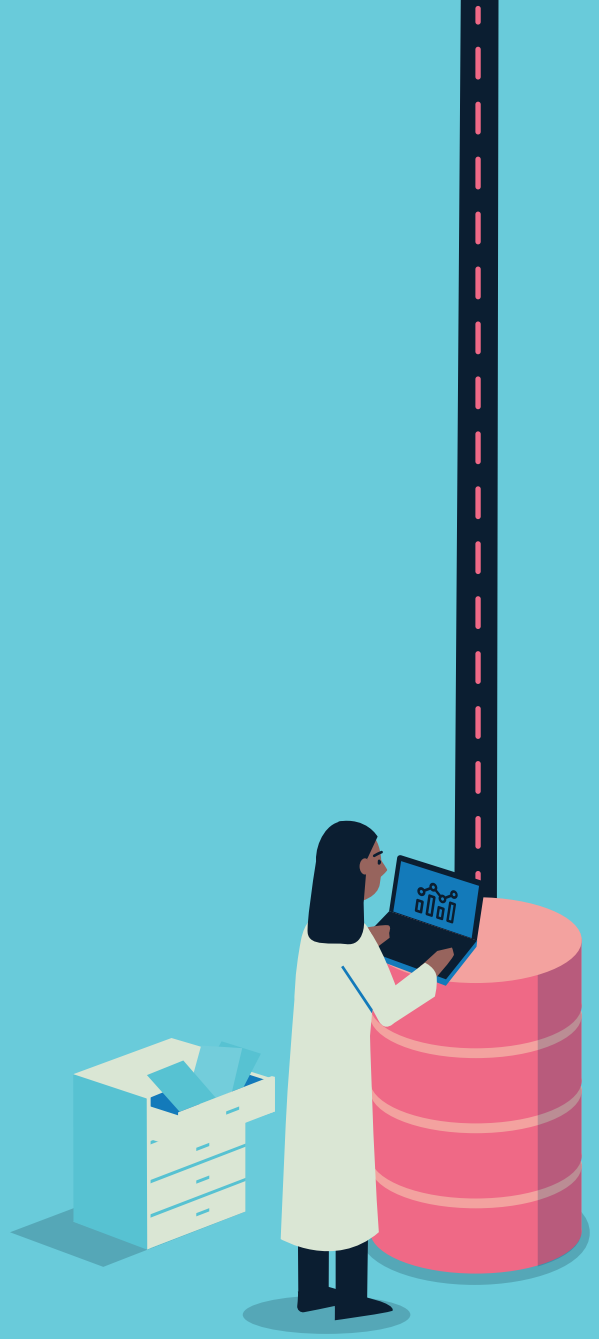
However, we see this as an exciting and worthwhile challenge. By making argument-based assurance FAIRer, we can help make the use of data-driven technologies more sustainable, accountable, explainable, and indeed, *fairer* for all.



Further Resources

Reports such as this one are static artefacts once published. However, the TEA platform's repository and documentation remains a living document. For up-to-date information about our current roadmap and focus areas, please visit one of the following:

- › GitHub repository (for roadmap and issues tracker):
<https://github.com/alan-turing-institute/AssurancePlatform>
- › Documentation site (for user guidance and additional training materials):
<https://alan-turing-institute.github.io/AssurancePlatform/>



References

- [1] E. G. Liberati et al., 'What can Safety Cases offer for patient safety? A multisite case study', *BMJ Qual. Saf.*, p. bmjqs-2023-016042, Sep. 2023, doi: [10.1136/bmjqs-2023-016042](https://doi.org/10.1136/bmjqs-2023-016042).
- [2] G. M. Cleland, I. Habli, J. Medhurst, and Health Foundation (Great Britain), *Evidence: using safety cases in industry and healthcare*. 2012.
- [3] UK Government, 'Ethnicity and COVID-19 from Race Disparity Unit'. Accessed: Jan. 19, 2024. [Online]. Available: <https://www.ethnicity-facts-figures.service.gov.uk/covid-19/>
- [4] A. Rajkomar, M. Hardt, M. D. Howell, G. Corrado, and M. H. Chin, 'Ensuring Fairness in Machine Learning to Advance Health Equity', *Ann. Intern. Med.*, vol. 169, no. 12, p. 866, Dec. 2018, doi: [10.7326/M18-1990](https://doi.org/10.7326/M18-1990).
- [5] T. Panch, H. Mattie, and R. Atun, 'Artificial intelligence and algorithmic bias: implications for health systems', *J. Glob. Health*, vol. 9, no. 2, p. 5, 2019, doi: [10.7189/jogh.09.020318](https://doi.org/10.7189/jogh.09.020318).
- [6] D. Leslie, A. Mazumder, A. Peppin, M. K. Wolters, and A. Hagerty, 'Does "AI" stand for augmenting inequality in the era of covid-19 healthcare?', *BMJ*, p. n304, Mar. 2021, doi: [10.1136/bmj.n304](https://doi.org/10.1136/bmj.n304).
- [7] P. Bachtiger, N. S. Peters, and S. L. Walsh, 'Machine learning for COVID-19—asking the right questions', *Lancet Digit. Health*, vol. 2, no. 8, pp. e391–e392, Aug. 2020, doi: [10.1016/S2589-7500\(20\)30162-X](https://doi.org/10.1016/S2589-7500(20)30162-X).
- [8] L. Sikstrom, M. M. Maslej, K. Hui, Z. Findlay, D. Z. Buchman, and S. L. Hill, 'Conceptualising fairness: three pillars for medical algorithms and health equity', *BMJ Health Care Inform.*, vol. 29, no. 1, p. e100459, Jan. 2022, doi: [10.1136/bmjhci-2021-100459](https://doi.org/10.1136/bmjhci-2021-100459).
- [9] J. W. Gichoya, L. G. McCoy, L. A. Celi, and M. Ghassemi, 'Equity in essence: a call for operationalising fairness in machine learning for healthcare', *BMJ Health Care Inform.*, vol. 28, no. 1, 2021, doi: [10.1136/bmjhci-2020-100289](https://doi.org/10.1136/bmjhci-2020-100289).
- [10] Department for Science, Innovation and Technology and Office for Artificial Intelligence, 'AI regulation: a pro-innovation approach', Aug. 2023. Accessed: Nov. 03, 2023. [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1176103/a-pro-innovation-approach-to-ai-regulation-amended-web-ready.pdf
- [11] Department for Science, Innovation and Technology, 'Introduction to AI assurance', Feb. 2024. Accessed: Mar. 08, 2024. [Online]. Available: https://assets.publishing.service.gov.uk/media/65ccf508c96cf3000c6a37a1/Introduction_to_AI_Assurance.pdf
- [12] European Commission, 'Regulatory framework proposal on artificial intelligence | Shaping Europe's digital future'. Accessed: Dec. 21, 2023. [Online]. Available: <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>
- [13] European Commission, 'Commission welcomes political agreement on AI Act', European Commission - European Commission. Accessed: Mar. 06, 2024. [Online]. Available: https://ec.europa.eu/commission/presscorner/detail/en/ip_23_6473
- [14] European Union, 'The AI Act Explorer | EU Artificial Intelligence Act'. Accessed: Mar. 15, 2024. [Online]. Available: <https://artificialintelligenceact.eu/ai-act-explorer/>
- [15] E. Tabassi, 'Artificial Intelligence Risk Management Framework (AI RMF 1.0)', National Institute of Standards and Technology (U.S.), Gaithersburg, MD, NIST AI 100-1, Jan. 2023, doi: [10.6028/NIST.AI.100-1](https://doi.org/10.6028/NIST.AI.100-1).
- [16] D. Elms, 'Limitations of risk approaches', *Civ. Eng. Environ. Syst.*, vol. 36, no. 1, pp. 2–16, Jan. 2019, doi: [10.1080/10286608.2019.1615474](https://doi.org/10.1080/10286608.2019.1615474).
- [17] I. Joshi and J. Morley, Eds., *Artificial Intelligence: How to get it right*. NHSX, 2019. Accessed: Nov. 01, 2023. [Online]. Available: [Trustworthy and Ethical Assurance of Digital Health and Healthcare](https://transform.</p></div><div data-bbox=)

[england.nhs.uk/media/documents/NHSX_AI_report.pdf](https://www.england.nhs.uk/media/documents/NHSX_AI_report.pdf)

- [18] K. Tobia, A. Nielsen, and A. Stremitzer, 'When Does Physician Use of AI Increase Liability?', *J. Nucl. Med.*, vol. 62, no. 1, pp. 17–21, Jan. 2021, [doi: 10.2967/jnumed.120.256032](https://doi.org/10.2967/jnumed.120.256032).
- [19] T. Lawton et al., 'Clinicians Risk Becoming "Liability Sinks" for Artificial Intelligence', *Authorea*, Jan. 2024, [doi: 10.22541/au.168209222.21704626](https://doi.org/10.22541/au.168209222.21704626).
- [20] M. Mitchell et al., 'Model Cards for Model Reporting', *Proc. Conf. Fairness Account. Transpar. - FAT 19*, pp. 220–229, 2019, [doi: 10.1145/3287560.3287596](https://doi.org/10.1145/3287560.3287596).
- [21] I. D. Raji, A. Smart, and R. N. White, 'Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing', p. 12, 2020.
- [22] D. Leslie et al., 'AI Ethics and Governance in Practice: An Introduction', The Alan Turing Institute., 2023. Accessed: Dec. 21, 2023. [Online]. Available: <https://www.turing.ac.uk/news/publications/ai-ethics-and-governance-practice-introduction>
- [23] Z. Porter, I. Habli, J. McDermid, and M. Kaas, 'A principles-based ethics assurance argument pattern for AI and autonomous systems', *AI Ethics*, Jun. 2023, [doi: 10.1007/s43681-023-00297-2](https://doi.org/10.1007/s43681-023-00297-2).
- [24] J. Mökander and L. Floridi, 'Ethics-Based Auditing to Develop Trustworthy AI', *Minds Mach.*, Feb. 2021, [doi: 10.1007/s11023-021-09557-8](https://doi.org/10.1007/s11023-021-09557-8).
- [25] D. Leslie, 'Understanding artificial intelligence ethics and safety', The Alan Turing Institute, Jun. 2019. Accessed: Mar. 25, 2020. [Online]. Available: https://www.turing.ac.uk/sites/default/files/2019-06/understanding_artificial_intelligence_ethics_and_safety.pdf
- [26] S. Toulmin, *The Uses of Argument*, Updated Edition. Cambridge: Cambridge University Press, 2003.
- [27] C. Burr and D. Leslie, 'Ethical assurance: a practical approach to the responsible design, development, and deployment of data-driven technologies', *AI Ethics*, Jun. 2022, [doi: 10.1007/s43681-022-00178-0](https://doi.org/10.1007/s43681-022-00178-0).
- [28] R. Bloomfield and P. Bishop, 'Safety and Assurance Cases: Past, Present and Possible Future – an Adelard Perspective', in *Making Systems Safer*, C. Dale and T. Anderson, Eds., London: Springer London, 2010, pp. 51–67. [doi: 10.1007/978-1-84996-086-1_4](https://doi.org/10.1007/978-1-84996-086-1_4).
- [29] I. Habli, S. White, M. Suján, S. Harrison, and M. Ugarte, 'What is the safety case for health IT? A study of assurance practices in England', *Saf. Sci.*, vol. 110, pp. 324–335, 2018, [doi: 10.1016/j.ssci.2018.09.001](https://doi.org/10.1016/j.ssci.2018.09.001).
- [30] M. Suján and I. Habli, 'Changing the patient safety mindset: can safety cases help?', *BMJ Qual. Saf.*, p. bmjqs-2023-016652, Nov. 2023, [doi: 10.1136/bmjqs-2023-016652](https://doi.org/10.1136/bmjqs-2023-016652).
- [31] C. Burr and R. Powell, 'Trustworthy Assurance of Digital Mental Healthcare', Alan Turing Institute, Nov. 2022. [doi: 10.5281/zenodo.7107200](https://doi.org/10.5281/zenodo.7107200).
- [32] 'AI Safety Summit', AISS 2023. Accessed: Nov. 03, 2023. [Online]. Available: <https://www.aisafetysummit.gov.uk/>
- [33] Ada Lovelace Institute, 'A knotted pipeline: data-driven systems and inequalities in health and social care', Nov. 2022. Accessed: Jan. 19, 2024. [Online]. Available: <https://www.adalovelace-institute.org/wp-content/uploads/2022/11/Ada-Lovelace-Institute-Knotted-pipeline-Data-driven-systems-and-inequalities-in-health-and-social-care.pdf>
- [34] The Health Foundation, 'Complex adaptive systems', Aug. 2010. [Online]. Available: <https://www.health.org.uk/publications/complex-adaptive-systems>
- [35] G. Sirugo, S. M. Williams, and S. A. Tishkoff, 'The Missing Diversity in Human Genetic Studies', *Cell*, vol. 177, no. 1, pp. 26–31, Mar. 2019, [doi: 10.1016/j.cell.2019.02.048](https://doi.org/10.1016/j.cell.2019.02.048).

- [36] C. R. Green et al., 'The Unequal Burden of Pain: Confronting Racial and Ethnic Disparities in Pain', *Pain Med.*, vol. 4, no. 3, pp. 277–294, Sep. 2003, doi: [10.1046/j.1526-4637.2003.03034.x](https://doi.org/10.1046/j.1526-4637.2003.03034.x).
- [37] M. E. Morales and R. J. Yong, 'Racial and Ethnic Disparities in the Treatment of Chronic Pain', *Pain Med.*, vol. 22, no. 1, pp. 75–90, Jan. 2021, doi: [10.1093/pm/pnaa427](https://doi.org/10.1093/pm/pnaa427).
- [38] D. Wen et al., 'Characteristics of publicly available skin cancer image datasets: a systematic review', *Lancet Digit. Health*, vol. 4, no. 1, pp. e64–e74, Jan. 2022, doi: [10.1016/S2589-7500\(21\)00252-1](https://doi.org/10.1016/S2589-7500(21)00252-1).
- [39] J. H. Kim and A. R. Scialli, 'Thalidomide: The Tragedy of Birth Defects and the Effective Treatment of Disease', *Toxicol. Sci.*, vol. 122, no. 1, pp. 1–6, Jul. 2011, doi: [10.1093/toxsci/kfr088](https://doi.org/10.1093/toxsci/kfr088).
- [40] 'Brazil's new generation of Thalidomide babies', *BBC News*, Jul. 24, 2013. Accessed: Dec. 15, 2023. [Online]. Available: <https://www.bbc.com/news/magazine-23418102>
- [41] M. De Martinis, M. M. Sirufo, M. Polsinelli, G. Placidi, D. Di Silvestre, and L. Ginaldi, 'Gender Differences in Osteoporosis: A Single-Center Observational Study', *World J. Mens Health*, vol. 39, no. 4, pp. 750–759, Oct. 2021, doi: [10.5534/wjmh.200099](https://doi.org/10.5534/wjmh.200099).
- [42] P. Gorczynski and F. Fasoli, 'LGBTQ+ focused mental health research strategy in response to COVID-19', *Lancet Psychiatry*, vol. 7, no. 8, p. e56, Aug. 2020, doi: [10.1016/S2215-0366\(20\)30300-X](https://doi.org/10.1016/S2215-0366(20)30300-X).
- [43] T. T. W. Community et al., 'The Turing Way: A Handbook for Reproducible Data Science'. Zenodo, Mar. 25, 2019. Accessed: Oct. 20, 2021. [Online]. Available: <https://the-turing-way.netlify.app/welcome>
- [44] L. Floridi and J. Cows, 'A Unified Framework of Five Principles for AI in Society', *Harv. Data Sci. Rev.*, Jun. 2019, doi: [10.1162/99608f92.8c-d550d1](https://doi.org/10.1162/99608f92.8c-d550d1).
- [45] Alan Turing Institute, 'Human Rights, Democracy, and the Rule of Law Assurance Framework for AI Systems: A Proposal', 2021.
- [46] P. Braveman, S. Egerter, and D. R. Williams, 'The Social Determinants of Health: Coming of Age', *Annu. Rev. Public Health*, vol. 32, no. 1, pp. 381–398, Apr. 2011, doi: [10.1146/annurev-publhealth-031210-101218](https://doi.org/10.1146/annurev-publhealth-031210-101218).
- [47] World Health Organization, 'Health equity and its determinants'. Accessed: Oct. 25, 2023. [Online]. Available: <https://www.who.int/publications/m/item/health-equity-and-its-determinants>
- [48] World Health Organization, 'Rio Political Declaration on Social Determinants of Health', presented at the World Conference on Social Determinants of Health, Rio de Janeiro: World Health Organization, 2011, p. 7. Accessed: Aug. 19, 2022. [Online]. Available: <https://www.who.int/publications/m/item/rio-political-declaration-on-social-determinants-of-health>
- [49] H. Graham, 'Social Determinants and Their Unequal Distribution: Clarifying Policy Understandings', *Milbank Q.*, vol. 82, no. 1, pp. 101–124, Mar. 2004, doi: [10.1111/j.0887-378X.2004.00303.x](https://doi.org/10.1111/j.0887-378X.2004.00303.x).
- [50] Public Health England, 'Health Profile for England 2021'. Accessed: Nov. 09, 2023. [Online]. Available: https://fingertips.phe.org.uk/static-reports/health-profile-for-england/hpfe_report.html
- [51] G. Dahlgren and M. Whitehead, 'The Dahlgren-Whitehead model of health determinants: 30 years on and still chasing rainbows', *Public Health*, vol. 199, pp. 20–24, Oct. 2021, doi: [10.1016/j.puhe.2021.08.009](https://doi.org/10.1016/j.puhe.2021.08.009).
- [52] C. E. Pollack, S. Chideya, C. Cubbin, B. Williams, M. Dekker, and P. Braveman, 'Should Health Studies Measure Wealth?', *Am. J. Prev. Med.*, vol. 33, no. 3, pp. 250–264, Sep. 2007, doi: [10.1016/j.amepre.2007.04.033](https://doi.org/10.1016/j.amepre.2007.04.033).
- [53] N. Daniels, B. P. Kennedy, and I. Kawachi, 'Why Justice Is Good for Our Health: The Social De-

terminants of Health Inequalities', *Daedalus*, vol. 128, no. 4, pp. 215–251, 1999.

- [54] D. Black and P. Townsend, Eds., *Inequalities in health: the black report*, Repr. in A Pelican original. Harmondsworth, Middlesex: Penguin Books, 1984.
- [55] M. G. Marmot, G. Rose, M. Shipley, and P. J. Hamilton, 'Employment grade and coronary heart disease in British civil servants.', *J. Epidemiol. Community Health*, vol. 32, no. 4, pp. 244–249, Dec. 1978, [doi: 10.1136/jech.32.4.244](https://doi.org/10.1136/jech.32.4.244).
- [56] M. G. Marmot, *The status syndrome: how your social standing affects our health and longevity*, 1st Holt paperbacks ed. New York, NY: Holt, 2005.
- [57] M. Grieves and J. Vickers, 'Digital Twin: Mitigating Unpredictable, Undesirable Emergent Behavior in Complex Systems', in *Transdisciplinary Perspectives on Complex Systems*, F.-J. Kahlen, S. Flumerfelt, and A. Alves, Eds., Cham: Springer International Publishing, 2017, pp. 85–113. [doi: 10.1007/978-3-319-38756-7_4](https://doi.org/10.1007/978-3-319-38756-7_4).
- [58] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi, 'Fairness and Abstraction in Sociotechnical Systems', in *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*, Atlanta, GA, USA: ACM Press, 2019, pp. 59–68. [doi: 10.1145/3287560.3287598](https://doi.org/10.1145/3287560.3287598).
- [59] ISO/IEC, 'ISO/IEC DIS 23894 Artificial intelligence — Risk management', British Standards Institution, 2022.
- [60] International Organization for Standardization, 'ISO 31000:2018 risk management – guidelines'. International Organization for Standardization, 2018. [Online]. Available: <https://www.iso.org/iso-31000-risk-management.html>
- [61] M. Schuijff and A. M. Dijkstra, 'Practices of Responsible Research and Innovation: A Review', *Sci. Eng. Ethics*, vol. 26, no. 2, pp. 533–574, Apr. 2020, [doi: 10.1007/s11948-019-00167-3](https://doi.org/10.1007/s11948-019-00167-3).
- [62] S. Agarwal and S. Mishra, *Responsible AI*. Springer, 2021.
- [63] N. D. Cara and N. Zelenka, 'Data Hazards', Jun. 2021, [doi: 10.17605/OSF.IO/3FV7T](https://doi.org/10.17605/OSF.IO/3FV7T).
- [64] IBM Research, 'Introducing AI Fairness 360, A Step Towards Trusted AI', *IBM Research Blog*. Accessed: Dec. 03, 2020. [Online]. Available: <https://www.ibm.com/blogs/research/2018/09/ai-fairness-360/>
- [65] B. Ozturk, T. Lawton, S. Smith, and I. Habli, 'Predicting Progression of Type 2 Diabetes Using Primary Care Data with the Help of Machine Learning', in *Studies in Health Technology and Informatics*, M. Hägglund, M. Blusi, S. Bonacina, L. Nilsson, I. Cort Madsen, S. Pelayo, A. Moen, A. Benis, L. Lindsköld, and P. Gallos, Eds., IOS Press, 2023. [doi: 10.3233/SHTI230060](https://doi.org/10.3233/SHTI230060).
- [66] A. Dinh, S. Miertschin, A. Young, and S. D. Mohanty, 'A data-driven approach to predicting diabetes and cardiovascular disease with machine learning', *BMC Med. Inform. Decis. Mak.*, vol. 19, no. 1, pp. 1–15, 2019.
- [67] P. Ryan Conmy, B. Ozturk, T. Lawton, and I. Habli, 'The Impact of Training Data Shortfalls on Safety of AI-Based Clinical Decision Support Systems', in *Computer Safety, Reliability, and Security*, vol. 14181, J. Guiochet, S. Tonetta, and F. Bitsch, Eds., in *Lecture Notes in Computer Science*, vol. 14181, Cham: Springer Nature Switzerland, 2023, pp. 213–226. [doi: 10.1007/978-3-031-40923-3_16](https://doi.org/10.1007/978-3-031-40923-3_16).
- [68] H. Carel and I. J. Kidd, 'Epistemic injustice in healthcare: a philosophical analysis', *Med. Health Care Philos.*, vol. 17, no. 4, pp. 529–540, Nov. 2014, [doi: 10.1007/s11019-014-9560-2](https://doi.org/10.1007/s11019-014-9560-2).
- [69] N. Égalité, 'The Epistemic Injustice of Racial Injustice', *Narrat. Inq. Bioeth.*, vol. 11, no. 3, pp. 259–264, Dec. 2021, [doi: 10.1353/nib.2021.0081](https://doi.org/10.1353/nib.2021.0081).
- [70] Y. Peled, 'Language barriers and epistemic injustice in healthcare settings', *Bioethics*, vol. 32, no. 6, pp. 360–367, Jul. 2018, [doi: 10.1111/bioe.12435](https://doi.org/10.1111/bioe.12435).

- [71] I. Contreras and J. Vehi, 'Artificial Intelligence for Diabetes Management and Decision Support: Literature Review', *J. Med. Internet Res.*, vol. 20, no. 5, p. e10775, May 2018, [doi: 10.2196/10775](https://doi.org/10.2196/10775).
- [72] Y. Jia, J. McDermid, T. Lawton, and I. Habli, 'The Role of Explainability in Assuring Safety of Machine Learning in Healthcare', *IEEE Trans. Emerg. Top. Comput.*, vol. 10, no. 4, pp. 1746–1760, Oct. 2022, [doi: 10.1109/TETC.2022.3171314](https://doi.org/10.1109/TETC.2022.3171314).
- [73] M. C. Elish, 'Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction', *Engag. Sci. Technol. Soc.*, vol. 5, pp. 40–60, Mar. 2019, [doi: 10.17351/ests2019.260](https://doi.org/10.17351/ests2019.260).
- [74] I. Sim et al., 'Reproducibility of Atrial Fibrosis Assessment Using CMR Imaging and an Open Source Platform', *JACC Cardiovasc. Imaging*, vol. 12, no. 10, pp. 2076–2077, Oct. 2019, [doi: 10.1016/j.jcmg.2019.03.027](https://doi.org/10.1016/j.jcmg.2019.03.027).
- [75] O. Razeghi et al., 'Fully Automatic Atrial Fibrosis Assessment Using a Multilabel Convolutional Neural Network', *Circ. Cardiovasc. Imaging*, vol. 13, no. 12, p. e011512, Dec. 2020, [doi: 10.1161/CIRCIMAGING.120.011512](https://doi.org/10.1161/CIRCIMAGING.120.011512).
- [76] L. C. Li, J. M. Grimshaw, C. Nielsen, M. Judd, P. C. Coyte, and I. D. Graham, 'Evolution of Wenger's concept of community of practice', *Implement. Sci.*, vol. 4, no. 1, p. 11, Dec. 2009, [doi: 10.1186/1748-5908-4-11](https://doi.org/10.1186/1748-5908-4-11).
- [77] NHS England, 'The NHS AI Lab—Accelerating the safe adoption of artificial intelligence in health and care', *NHS Transformation Directorate*. Accessed: Jan. 15, 2024. [Online]. Available: <https://transform.england.nhs.uk/ai-lab/>
- [78] D. Leslie et al., 'AI Ethics and Governance in Practice: AI Fairness in Practice', Alan Turing Institute, 2023. Accessed: Jan. 12, 2024. [Online]. Available: <https://www.turing.ac.uk/news/publications/ai-ethics-and-governance-practice-ai-fairness-practice>
- [79] The Assurance Case Working Group, 'Assurance Case Guidance: Challenges, Common Issues and Good Practice', Aug. 2021. Accessed: Nov. 10, 2023. [Online]. Available: <https://scsc.uk/r159:1>
- [80] The Assurance Case Working Group, 'GSN Community Standard Version 3', May 2021. [Online]. Available: <https://scsc.uk/r141C:1?t=1>
- [81] C. Burr and D. Leslie, 'Ethical Assurance: A practical approach to the responsible design, development, and deployment of data-driven technologies'. 2021.
- [82] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian, 'The (Im)possibility of fairness: different value systems require different mechanisms for fair decision making', *Commun. ACM*, vol. 64, no. 4, pp. 136–143, Apr. 2021, [doi: 10.1145/3433949](https://doi.org/10.1145/3433949).
- [83] B. Ruf and M. Detyniecki, 'Towards the Right Kind of Fairness in AI', AXA, Oct. 2021. Accessed: Nov. 24, 2023. [Online]. Available: https://www-axa-com.cdn.axa-contento-118412.eu/www-axa-com/d6324958-367e-4375-81c3-cfeb8e7ccc66_AXA_FairnessCompass-English.pdf
- [84] T. Gebru et al., 'Datasheets for Datasets', in *Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning*, 2018. Accessed: Apr. 06, 2020. [Online]. Available: <http://arxiv.org/abs/1803.09010>
- [85] T. Scanlon, *What we owe to each other*. Cambridge (Mass.): the Belknap press of Harvard University Press, 2000.
- [86] Health and Safety Executive, 'Risk management: Expert guidance - ALARP at a glance'. Accessed: Mar. 15, 2024. [Online]. Available: <https://www.hse.gov.uk/enforce/expert/alarpglance.htm>
- [87] D. B. Deutz et al., 'How to FAIR: a website to guide researchers on making research data more FAIR'. Zenodo, Jun. 2020. [doi: 10.5281/zenodo.3712065](https://doi.org/10.5281/zenodo.3712065).
- [88] DSW, 'Data Stewardship Wizard', Data Stewardship Wizard. Accessed: Jan. 02, 2024. [Online]. Available: <https://ds-wizard.org/>

- [89] M. Thompson, K. Burger, R. Kaliyaperumal, M. Roos, and L. O. B. Da Silva Santos, 'Making FAIR Easy with FAIR Tools: From Creolization to Convergence', *Data Intell.*, vol. 2, no. 1–2, pp. 87–95, Jan. 2020, [doi: 10.1162/dint_a_00031](https://doi.org/10.1162/dint_a_00031).
- [90] J. Rawls, *Justice as fairness: A restatement*. Harvard University Press, United States, 2001.
- [91] A. Sen, *The Idea of Justice*. Penguin, United Kingdom, 2010.



Trustworthy and Ethical Assurance of Digital Health and Healthcare (TEA-DH) Project and Engagement



About the appendix

This section provides a summary of the TEA-DH project, focusing on the project's objectives and the stakeholder engagement work that was carried out to support these objectives. Themes were identified for each of the workshops we organised, as well as cross-cutting themes.

This report was produced as part of the Trustworthy and Ethical Assurance of Digital Health and Healthcare (TEA-DH) project, which was generously funded by the Assuring Autonomy International Programme, a partnership between Lloyd's Register Foundation and the University of York.

The project commenced in April 2023 and ran until December 2023 and had the following three objectives:

- 1. Develop the existing TEA methodology and platform:** work had already been undertaken to explore how the methodology and platform could enable a more systematic and accessible approach to the assurance of trustworthy and ethical goals, but in the context of digital mental healthcare. As such, this project was building upon existing foundations but expanding our work to a broader domain (i.e. digital health and healthcare).
- 2. Improve impact:** the trustworthy and ethical assurance methodology and platform had already been evaluated and tested with diverse stakeholder groups. This project sou-

ght to continue developing their potential impact by doing further engagement with regulators and policymakers, as part of a wider push to enhance the UK's AI assurance ecosystem and address core needs and capabilities gaps.

- 3. User experience enhancements and validation:** it is not enough to just develop a methodology or tool to address key needs, they also must be usable. This means investing in UI/UX.

To ensure these objectives were met, the project carried out a range of engagement workshops with the aim of ensuring that our work was addressing actual needs and challenges, as well as building impact in the process.

Over the course of several months, we ran workshops with the following groups:

- › **Regulators and policymakers** (including representatives from the Care Quality Commission; the Equality and Human Rights Commission; the Information Commissioner's Office; NHS England; the Department for Science Innovation, and Technology; and the Law Society). The following questions were posed to the group:
 - › What role would regulators and policymakers like to see themselves playing in assessing and evaluating arguments and evidence that an AI system or data-driven technology is fair?
 - › At what stage of a technology's lifecycle do regulators and policymakers see their involvement in assessing/evaluating fairness of an AI system?
 - › What responsibilities and duties do regulators currently have and see themselves as having to safeguard fairness and equity?
 - › What powers do regulators and policymakers have to ensure service providers uphold and ensure fairness and health equity?
 - › What are the obstacles faced by regulators and policymakers when evaluating digital healthcare systems for fairness and approving them?
 - › What is the expected or anticipated worth of an assurance argument in the area of fairness and digital health and healthcare?
- › **Practitioners** (including developers and software engineers, business and product managers, and other senior decision-makers from organisations building data-driven technologies). The following questions were posed to the group:
 - › Who is most likely to make decisions about ethical practices in a company adopting AI systems?
 - › Who else would need to be involved to support the adherence of these principles?
 - › Whose responsibility would it be to monitor these actions/ethical principles being adhered to?

- › What are the main three challenges companies face when trying to adhere to ethical principles?
- › What are organisations trying to do to mitigate these challenges?
- › What are the risks for companies not adopting these principles?
- › What do companies currently know about ethical practices in AI systems?
- › How are companies currently adhering to AI principles?
- › What tools or resources are being used to support people to meet AI principles?
- › What knowledge gaps exist in teams creating ethical principles within AI systems?
- › What motivates a company to start thinking about adopting more ethical practices?
- › At what point in the product life cycle do ethical considerations take place?
- › What triggers teams to put these practices in place?
- › What benefits beyond safety and reducing risk can be achieved by adhering to ethical practices for the businesses?
- › In addition, the participants were also asked to construct a partial assurance case for a case study exploring the fairness of a hypothetical AI system in healthcare.
- › **Researchers** (from disciplines including data science, medicine and public health, law, philosophy, and social sciences). Unlike the other groups, the researchers were asked to identify specific claims and possible forms of evidence that could be used to provide assurance for the case study presented in **Section 4** (diabetes risk prediction) and to identify a specific stage of the project lifecycle model (see **Figure 2.5**) where these claims were likely to be located.

As the engagement activities were designed solely to support the three objectives above, we did not set out to publish a thematic review of the workshop data. Nevertheless, we did analyse and explore the findings to see where common themes emerged. This analysis was carried out by three members of the project team, who did initial, independent coding before reaching consensus on the following themes.

Workshop 1 (Regulators and Policy-Makers)

- › **Importance of assurance ecosystem context:** it was recognised that there are many techniques, tools, processes, frameworks, and standards for assurance. As such, it is important to be aware of how assurance is situated within these, often overlapping, contexts. A pluralistic approach to understanding assurance should be pursued, while

recognising the need to have a shared vocabulary for the assurance ecosystem (e.g. roles and responsibilities).

- › **Gaps in skills and capabilities:** to ensure the Trustworthy and Ethical Assurance platform is successful and functions effectively, skills and capability building will be crucial. However, there are currently many gaps in skills and general capabilities (e.g., between intentions and outcomes, between a developer's knowledge/understanding and a regulator's knowledge/understanding, between when regulators can intervene in the AI lifecycle and when intervention ought to occur in the AI lifecycle, etc.). Case studies that demonstrate how assurance goals, such as *fairness*, have been operationalised throughout a project's lifecycle are one way that these gaps can be addressed, but additional capability building will be needed.
- › **Distributed responsibilities:** in addition to being one piece of the assurance ecosystem, the responsibility of regulators to ensure safe and fair deployment and use of data-driven technologies is distributed across multiple regulators. This makes the process of assurance and verification difficult (e.g. verifying claims and evidence, enforcing compliance, monitoring ongoing impact). This is particularly important in the context of assurance of fairness, as the equitable impact of a system will (in part) depend on how a system is used. Using assurance cases as a means for requiring developers of technologies to proactively explain and justify how they have made a system that is fair, for example, was seen as a positive shift in the burden of responsibility. However, as noted above, this is not without its challenges (e.g. how can a developer assure fair use of their system post-deployment?), and regulatory gaps will still need to be addressed.

Workshop 2 (Practitioners)

- › **Importance of assurance ecosystem context:** as with the previous workshop, context was seen as vital to understanding both the benefits and challenges of assurance. However, the understanding of this context was framed around topics such as the importance of liability for organisations and how assurance can help limit and mitigate risks. Reliance on regulatory guidance, legislation, and internal mechanisms (e.g. red teaming or advisory boards), reinforced the need to think about the assurance ecosystem as comprising a diverse set of roles and responsibilities.
- › **Distributed responsibilities:** whereas regulators have varying remits of responsibility (e.g. regulation of devices versus services), practitioners also have distributed responsibilities across the project's lifecycle of system. This was recognised as a challenge to the creation of a sufficient and comprehensive form of assurance for ethical goals, which by their very nature are typically wide-ranging in

scope. However, having diverse and multi-disciplinary conversations (e.g. inclusive stakeholder engagement) was understood as an important priority by many.

- › **Gaps in skills and capabilities:** as above, practitioners highlighted the challenge of skills and capabilities gaps. However, two additional and more specific challenges emerged in relation to this theme. First, practitioners noted that the gaps are further exacerbated by *limited resources, skills, and motivation*. For instance, ethical reflection or decision-making is typically triggered only when specific risks are identified or when harms emerge and there are concerns voiced about possible reputational damage. As such, a more anticipatory approach to ethical reflection and decision-making is often given low priority due to limited resources. However, where there are strong legislative or regulatory requirements, such as enforcement of security and data privacy, more resources have led to ways of identifying and addressing related risks or harms. Second, a *lack of standardisation and consensus* around ethical principles was viewed by some as a reason for low adoption or development of best practices.²³

Workshop 3 (Researchers)

The workshop with researchers was conducted with a different case study and activity. As such, the themes diverge from the first two workshops.

- › **Appropriateness of assurance claims and evidence:** although bearing some resemblance to the above theme of 'assurance ecosystem context', the notion of 'appropriateness' is central to this theme. Several participants, in different groups, identified that what will be appropriate for one stakeholder regarding the sufficiency or justifiability of a fairness case will not be for another. For instance, whereas a specific form of evidence may suffice to warrant an associated fairness claim in one context, there is no guarantee that this will hold for other use contexts. This may be due to varying norms or expectations from stakeholders (e.g. different fairness requirements or cultural expectations), or because of different regulatory requirements (e.g. public versus private sector).
- › **Diversity in team composition:** the need to have a diverse project team, or to carry out inclusive forms of stakeholder engagement and meaningful participatory design, was seen by many as an important precondition for establishing a convincing and justifiable assurance case. In some cases, a specific claim about the project team and the inclusivity of the project's governance was seen as necessary for a fairness case.

²³ This has also been highlighted by others in published research focused on addressing algorithmic bias in the context of AI and health systems [5].

- › **Challenges of identifying and formulating fairness claims:** related to the 'gaps in skills and capabilities' theme from the previous workshop, this theme recognises a difficulty that several participants found when developing claims that could be considered *claims about fairness* in their own right, rather than just good practices or pre-requisites for good project governance and management (e.g. accountability of project leads).
- › **Practical tools guardrails:** in constructing their hypothetical assurance case, many participants recognised the importance of establishing pragmatic mechanisms that could be seen as opening the assurance of fairness claims to independent assessment, evaluation, scrutiny, or oversight (e.g. review board milestones and checkpoints; transparent and accessible bias audits and fairness requirements). This also connects with the prior theme related to ensuring a flourishing assurance ecosystem.

Cross-Cutting Themes

From the above, it seems fair to conclude that the following cross-cutting themes emerged across the groups, but retained important nuances unique to the groups:

- › Awareness of context is vital when carrying out assurance activities, especially when considering goals such as fairness. Here, context may refer to the broader regulatory context, the research and development context, the use context, or the ethical context, among other things.
- › There is a pressing need to cultivate a flourishing assurance ecosystem, in which many different techniques, tools, processes and standards are available. However, a shared vocabulary—especially around common goals, values, or objectives—should underpin this plurality.
- › There are many challenges posed by current gaps in skills and capabilities, which will create barriers to the adoption of responsible research and innovation practices. Organisations should continue to invest resources and build capabilities through upskilling activities and strategic planning.

The Trustworthy and Ethical Assurance (TEA) Platform



About the appendix

This section presents technical details of the TEA platform, including a selection of key features that were either revised or developed based on engagement and/or UX design work carried out during the TEA-DH project.

The Trustworthy and Ethical Assurance (TEA) platform, developed by researchers at The Alan Turing Institute and University of York, is an innovative, open-source tool designed to facilitate the process of creating, managing, and sharing assurance cases for data-driven technologies, such as digital twins or AI.

Our goal is to make the complex world of assurance accessible to a broad range of stakeholders and affected users, including researchers, developers, auditors, regulators, and decision-makers.

With our interactive tools, comprehensive educational resources, and a supportive community infrastructure, the TEA platform can aid you in simplifying the development of arguments and evidence for goals such as safety and sustainability, accountability, fairness, and explainability. By streamlining the process of developing and communicating a structured assurance case, the TEA platform fosters a more inclusive ecosystem of trustworthy and ethical technology governance.

Technical Details

Our technology stack ensures that the TEA platform is not only powerful and reliable but also accessible to users with different levels of technical expertise.

At its core, the platform features a web application built with the [React](#) framework, known for its modular and interactive user interfaces. This is complemented by use of the [Material UI](#) (MUI) component library, which enables straightforward and consistent adoption of accessible and intuitive design elements. The web application also uses the open-source package, [Mermaid.js](#), to transform complex assurance cases into understandable flowcharts, enhancing user experience.

On the backend, the TEA platform is powered by [Django](#), a high-level Python web framework that offers robust backend capabilities, including a straightforward API for data management. Data can be stored in [SQLite](#) or [PostgreSQL](#) databases, providing options for lightweight to more scalable storage solutions.

The platform also supports easy installation and deployment through [Docker](#), making it straightforward to set up in various environments. Please visit our [documentation site](#) for further details on any of the above (e.g. specific details about the platform's API).

Key Features

1. Assurance Case Management

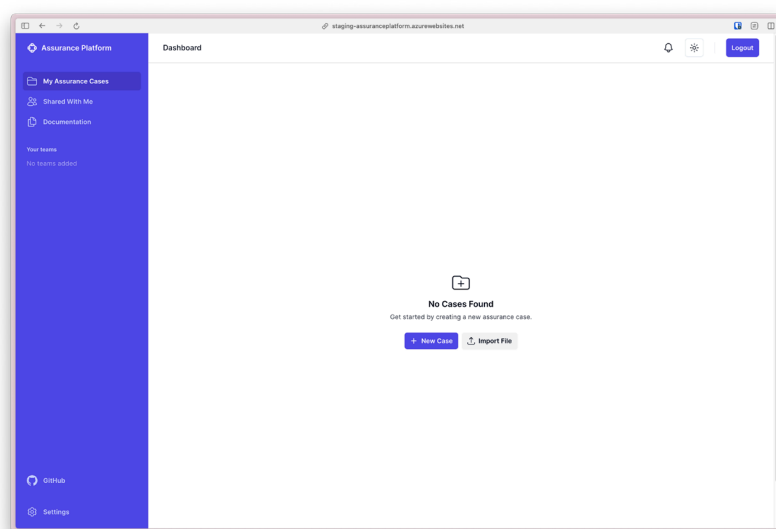


Figure A2.1: A screenshot of the TEA platform's assurance case management interface.

When logged in, users are greeted by a dashboard that exemplifies simplicity, initially displaying a prominent “Create a new case” button. This central feature is complemented by convenient access to the platform’s documentation and assurance methodology information on the left-hand side. Moreover, the dashboard offers functionality for importing existing assurance cases via the upper right corner button, showcasing the platform’s flexibility and user control.

2. Importing an Assurance Case

The TEA platform accommodates users’ needs for versatility by supporting file imports, specifically SVG or JSON formats which adhere to predefined conventions. This feature is detailed within our documentation site. It is our aim that users can seamlessly integrate their existing assurance cases into the platform, fostering a bridge between prior work and new collaborative opportunities. Enhancing the import feature, the platform also allows users to import assurance cases via URLs, accepting any publicly accessible link to SVG or JSON files that follow the platform’s file conventions. This addition broadens the scope for sharing and collaborating on assurance cases, making the process more inclusive and adaptable to various user needs. This will also feed into future plans to further develop upon our existing GitHub support (e.g. version control of assurance cases, OAuth), which will help users integrate the TEA platform into existing workflows.

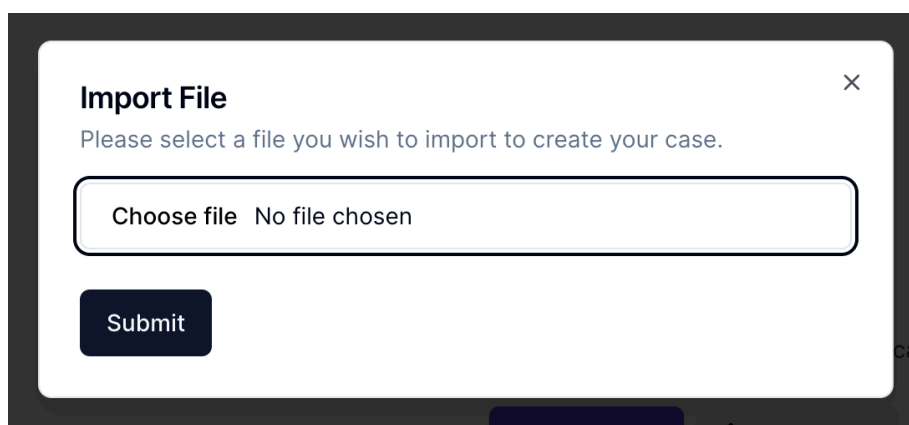


Figure A2.2: A screenshot of the 'Import File' modal.

3. Assurance Case Builder

The assurance case builder is the primary interface for the TEA platform. If a user chooses to start from scratch, they will encounter an empty but expansive interface. This has been designed to allow the user to focus on the assurance case itself, while providing easy access to the necessary tools and options. Following expert UI/UX guidance, subsequent actions are then easy to identify. For instance, adding a top-level goal is immediately identifiable from the button in the upper right corner, setting the foundation for the assurance case. Subsequent steps allow for the incorporation of the elements that are relevant to where the user is in the process of development. For instance, if a user has a property claims elements selected, they will be presented with two options ('Add Claim', 'Add Evidence'). However, the importing of argument patterns is also supported. For example, if the user has instead selected the 'Minimal' template, they will be presented with a pre-organised view featuring a Goal, its Context, and a Claim. This setup serves as an intuitive starting point, guiding users through the initial phases of assurance case development with a clear, manageable structure.

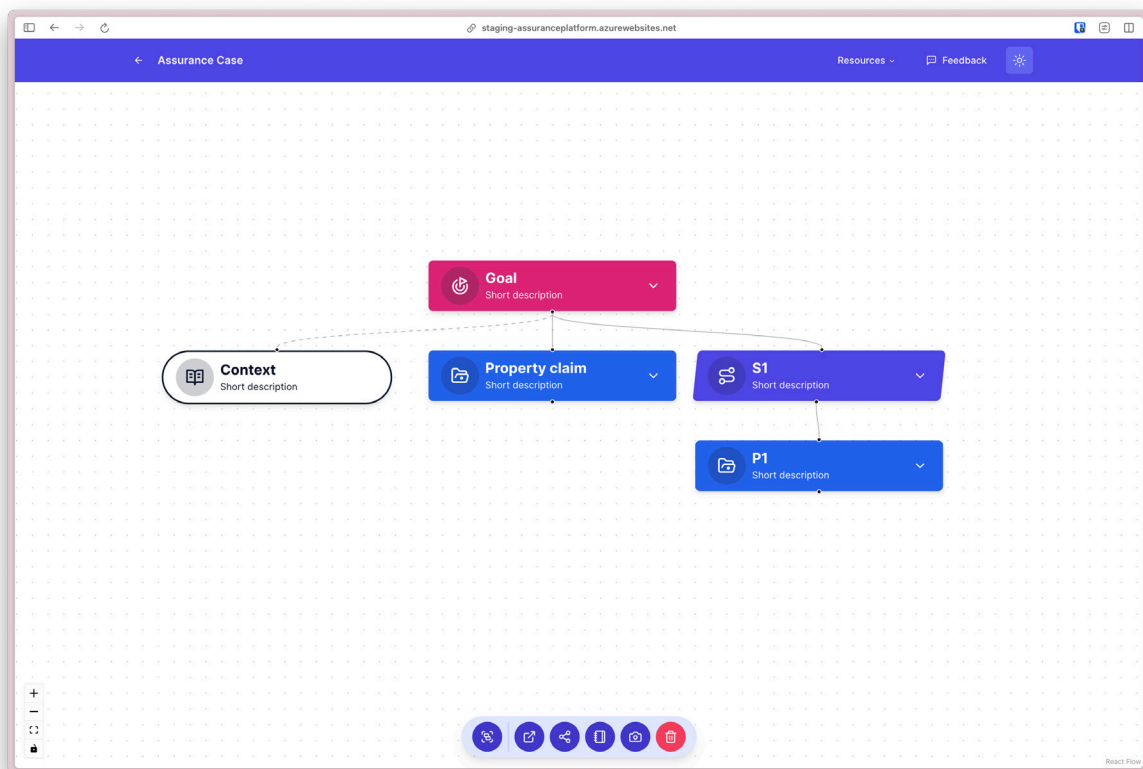


Figure A2.3: A screenshot of the assurance case builder (primary interface for the TEA platform).

4. Accessibility

We want the TEA platform to promote best practices for accessibility. While there is a lot to do here, we have started by considering visual presentation. For instance, in the Accessibility dialog, users can select from multiple colour schemes to customise the visual presentation of their assurance case, accommodating diverse user needs and preferences.

5. User Guidance and Documentation

To ensure the TEA platform is not just usable but also used, it is important that we provide clear user guidance and documentation that can help bring skills and training to a new and diverse audience. As such, we have created several introductory guides on Trustworthy and Ethical assurance, all of which are available through our project pages. We will continue to update these as the project develops and new features are added.

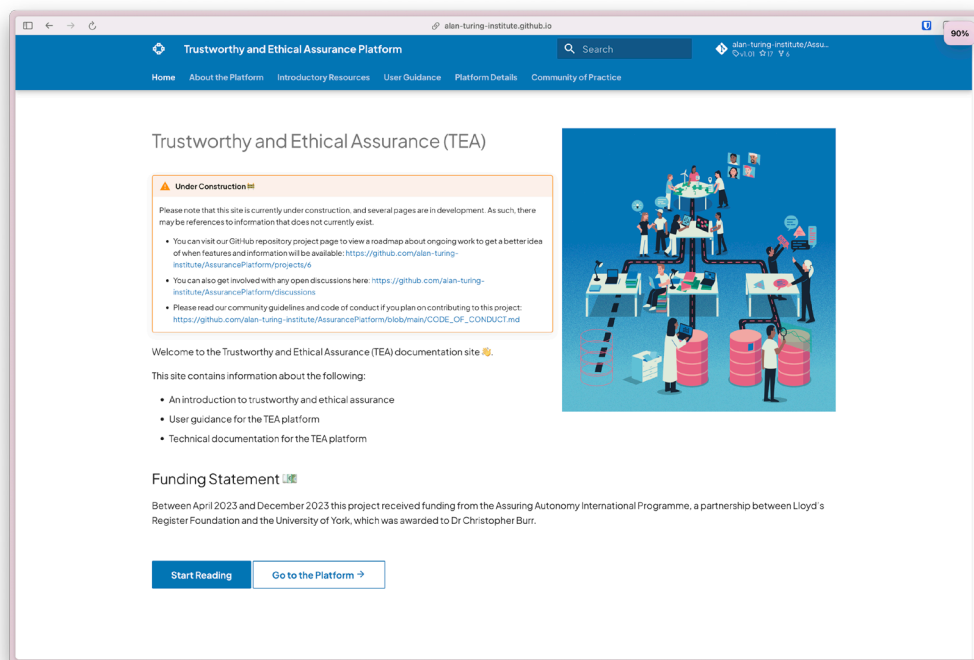


Figure A2.7: A screenshot of the user guidance and documentation site for the TEA platform.



**The
Alan Turing
Institute**

**turing.ac.uk
@turinginst**