


The Alan Turing Institute

Trustworthy Assurance of Digital Mental Healthcare

Dr Christopher Burr
Rosamund Powell



This report was authored and produced by
Dr Christopher Burr & Rosamund Powell
(cburr@turing.ac.uk, rpowell@turing.ac.uk)
(Public Policy Programme, The Alan Turing Institute).

Illustrations by Jonny Lighthands: www.jonnylighthands.co.uk
Graphic Design by Nerea Gómez: www.nereagomez.com
& Joudy Bourghli

Research and production for this report was supported by
funding from the UKRI's Trustworthy Autonomous Hub,
which was awarded to Dr Christopher Burr
(Grant number: TAS_PP_00040).

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

Burr, C. and Powell, R. (2022)
Trustworthy Assurance of Digital Mental Healthcare.
Public Policy Programme, The Alan Turing Institute.
<https://doi.org/10.5281/zenodo.7107200>

An online website also accompanies this report,
enabling accessibility features and providing additional information
that is not included in this publication:

<https://alan-turing-institute.github.io/trustworthy-assurance/>

**The
Alan Turing
Institute**



UKRI
**Trustworthy
Autonomous
Systems Hub**

Acknowledgements

Talking about mental health can be very difficult. It was important that we enabled and supported participants to share their perspectives and attitudes anonymously during the project's engagement events, and so we would like to begin by expressing our gratitude to those who took part in our events, or supported specific activities.

The research team would like to acknowledge the following groups and individuals:

- We are grateful for the candid and thoughtful discussions of the users of digital mental health technologies, who took part in our workshops facilitated by the McPin Foundation.
- The following individuals who generously gave their time to review a draft version of this report (alphabetical order): Dr Cath Biddle, Claudia Corradi, Dr Ibrahim Habli, Dr Zoe Porter, and Dr Mat Rawsthorne. We would also like to thank two anonymous reviewers.
- Members of the wider project team who offered helpful advice and guidance: Dr Kate Devlin, Professor David Leslie, and Morgan Briggs
- Our collaborators at University of York's Assuring Autonomy International Programme, including Dr Ibrahim Habli and Dr Zoe Porter who have been instrumental in ongoing discussions about argument-based assurance. And, also our thanks to all participants and the organisers of the **UKRI TAS Health and Social Care Workshop** for their valuable feedback and queries.
- Our collaborators at the McPin Foundation, including Dr Dan Robotham and Roya Camvar who were instrumental in facilitating conversations with individuals who have lived experiences relating to the use of digital mental health technologies.
- Colleagues at The Alan Turing Institute, who have provided direct and indirect input into this research project: Oana Romocea, Eirini Koutsouroupa, Pauline Kinniburgh, Dr Shyam Krishna, and Dr Michael Katell.
- With special thanks to all participants in our research engagements throughout this project, including participants with lived experiences of digital mental health technologies, students and administrators at UK universities and, finally, policymaker, researchers and developers working within the sector.

Table of Contents

Foreword	05
Executive Summary	07

Chapter 1	Introduction	> Laying the Foundations	16
		> The Current (Socioeconomic) Landscape of Digital Mental Health	19
		> A Culture of Distrust	24
		> About the Project	27

Chapter 2	Presenting Trustworthy Assurance—A Framework and Methodology	> Designing, Developing, and Deploying Trustworthy Digital Mental Health Technologies	33
		> What is Trustworthy Assurance	37
		> Argument Patterns	45

Chapter 3	Applying Trustworthy Assurance—Digital Mental Healthcare at UK Universities	> The University Context	54
		> Workshop Information	57
		> Analysis	59

Chapter 4	Co-Designing Trustworthy Assurance—Stakeholder Engagement	> Workshop Information	78
		> Analysis	81

Chapter 5	Developing Trustworthy Assurance—Argument Patterns for fairness and Explainability	> Co-designing argument patterns	109
------------------	---	----------------------------------	-----

Conclusion	126
Endnotes	130



Dr Cath Biddle
Head of Digital, Mind

In a society where mental health need far outstrips the supply of services, it is perhaps inevitable that policy-makers and regulators, healthcare professionals and service providers, and employers and educators are looking for scalable solutions. At the same time as digital and data-driven technologies have become mainstream in so many areas of our lives, from banking to dating, people seeking support are also turning to the internet and app stores to find ways to get help. And commercial organisations haven't been slow to spot this trend and provide solutions in exchange for your money or your data. For example, Calm and Headspace Health—both leading providers of mindfulness apps—were valued at \$2bn and \$3bn, and made an estimated \$200m and \$150m respectively in 2020.¹

In a fast moving and increasingly crowded marketplace, there is an urgent need to help those purchasing or procuring digital mental health services, for themselves or on behalf of others, to be confident that the products they are buying are not only safe and clinically effective, but also promote key ethical values, such as data protection, health equity and sustainability. In other words, are trustworthy.

At Mind, we are actively exploring how to use digital technology to increase the reach and impact of all aspects of our work, from fundraising to campaigning and service delivery. While we recognise that digital mental health services can increase choice and reach, they are not a panacea and cannot replace the localised, personal touch that is core to our service. **Our survey in 2021** of almost 2000 people revealed that more than one in three (35 per cent) found support from NHS mental health services, given over the phone or online, difficult to use; almost two in three (63 per cent) said they would have preferred to have been given face-to-face support; and one in four (23 per cent) say their mental health actually got worse as a result of using this support. However, two in three (69 per cent) appreciated not having to travel; almost one in two (47 per cent) were grateful for greater flexibility over appointment times; and two in five (40 per cent) said that waiting times were shorter. It is within this context that we are taking a test-and-learn approach to understand when and how digital mental health technologies can be used to augment and complement in person support.

As such, we welcome the work The Alan Turing Institute are doing to provide a framework to determine which digital mental health technologies are trustworthy.

Existing regulatory frameworks such as DTAC, NICE, MHRA, CQC exist, but do not provide sufficient coverage. Kooth, Togetherall and Silvercloud, for example, are all reputable, large companies delivering services to the NHS that fall outside the scope for CQC registration because the activities of those organisations are not deemed a 'regulated activity'. Many health-care apps also fall outside of the current definition of medical devices, and as such are outside of the scope of MHRA. And although NICE have recently updated their evidence standards framework for digital health technologies, it is not clear how evaluations will keep pace with the regularity of app updates.

This report raises questions that continue to concern providers of mental health support:

- › Can digital healthcare provide a route to support people who are not already in contact or not well served by formal healthcare services?
- › How can we ensure access to those who are digitally excluded or in data poverty, who may also be socially excluded and at increased risk of mental health challenges?
- › Can structured scrutiny, as described/proposed in this report, encourage developers to improve the transparency and trustworthiness of their products?
- › Are users (particularly at a time of increased vulnerability) adequately informed and protected by the practice of 'opting in' to terms and conditions which are often pages long and written for lawyers not end users?
- › Can digital mental healthcare replace the 'human connection' or is it safest and most effective when used to supplement and complement in person support?

The research and trustworthy assurance framework proposed in this report by The Alan Turing Institute's Public Policy Programme provides a key stepping stone towards addressing these questions, as we strive to meet the growing need for mental health support in a way that is both responsible and trustworthy.

Dr Cath Biddle

Head of Digital, Mind





There is a culture of distrust surrounding the development and use of digital mental health technologies (DMHTs).

As many organisations continue to grapple with the long-term impacts on mental health and well-being from the COVID-19 pandemic, a growing number are turning to digital technologies to increase their capacity and try to meet the growing need for mental health services. Prior to the pandemic, we had already called for greater attention to the ethical challenges of using digital technologies in the domain of psychiatry or mental healthcare.³ Since then, the urgency for meeting this call has only grown.

In this report, we argue that clearer assurance for how ethical principles have been considered and implemented in the design, development, and deployment of DMHTs is necessary to help build a more trustworthy and responsible ecosystem. To address this need, we set out a positive proposal for a framework and methodology we call 'Trustworthy Assurance'.

To support the development and evaluation of Trustworthy Assurance, we conducted a series of participatory stakeholder engagement events with students, University administrators, regulators and policy-makers, developers, researchers, and users of DMHTs. Our objectives were

- › to identify and explore how stakeholders understood and interpreted relevant ethical objectives for DMHTs,
- › to evaluate and co-design the trustworthy assurance framework and methodology, and
- › solicit feedback on the possible reasons for distrust in digital mental healthcare.

Based on these objectives, the following 'key findings' and 'recommendations' are presented.

Key Findings

- 1** The current landscape of digital mental healthcare is characterised by significant uncertainty, a lack of transparency or accountability, and a rising demand that outpaces trusted services and resources. This contributes to a culture of distrust, which may prevent vulnerable users from accessing support.

- 2** Concerns raised by stakeholders suggest that there are a wide range of challenges to be addressed, which may broadly be grouped into concerns surrounding a dearth of trustworthy innovation and concerns surrounding a lack of transparent communication between groups:
 - a.** For developers, key concerns focused on
 - › the lack of clear guidance and structure through which to present evidence of trustworthy innovation,
 - › the lack of integration of ethics within existing workflows, and
 - › the challenges posed by what is viewed as burdensome regulation.
 - b.** For policy-maker, key concerns focused on
 - › the lack of clarity surrounding standards for medical devices versus services for “well-being” that are widely available on digital platforms (e.g. app stores),
 - › the lack of integration or harmonisation between existing examples of legislation and standards in this space.
 - c.** For those with lived experiences of using these tools, key concerns focused on
 - › the lack of clear and meaningful consent procedures and the insufficiency of data privacy policies,
 - › the perceived erosion of in-person care by digital technologies and services,
 - › a perceived lack of diversity and representation in development teams,
 - › the varying quality and accessibility of services across society (e.g. the digital divide).

- 3** Trustworthy Assurance is a framework and methodology that can support the design, development, and deployment of data-driven technologies and also create a more responsible and trustworthy ecosystem of digital mental healthcare.

Recommendations

- 1** Organisations that are involved in the design, development, and deployment of DMHTs should adopt and use the trustworthy assurance methodology to demonstrate and justify how they have embedded core ethical principles into their systems. In doing so, the methodology can also help provide assurance for how key legislative or regulatory duties and obligations have been met.

- 2** Standards can be co-developed within and among organisations by sharing best practices related to trustworthy assurance. This can help ease the burden associated with relevant responsibilities (e.g. compliance, deliberation).

- 3** Common capacities should be developed across the digital mental healthcare landscape, such as initiatives aimed at improving data and digital literacy, in order to support and foster trustworthy and responsible innovation through shared best practices and standards.

- 4** Research should be undertaken to identify how organisations and product managers could ease the time burden on developers through embedding and integrating the trustworthy assurance methodology into key stages of the project lifecycle, rather than the methodology being treated as a post hoc compliance exercise.

- 5** Organisations involved in the design, development, or deployment of DMHTs should identify opportunities and processes to support the transformative and inclusive engagement and participation of affected users within the project lifecycle.

In each subsequent chapter, these key findings of our project are further contextualised and refined with reference to the specific topics under discussion.

Report Overview

Our report is structured as follows:

- **Chapter 1 (Introduction)** establishes the background context and conceptual foundations for the report, while also outlining the many challenges that exist for researchers, developers, and policy-makers/regulators working in the domain of digital mental healthcare.
- **Chapter 2 (Presenting Trustworthy Assurance)** introduces the framework and methodology of 'Trustworthy Assurance'. The framework includes a model of a typical project lifecycle involving a data-driven technology (e.g. health and well-being app), and a discussion of several ethical principles, known as the SAFE-D principles. The framework serves as a guide to our methodology for developing an assurance case that promotes trustworthy goals associated with DMHTs. Finally, this chapter also includes an important discussion about 'argument patterns', which supports the material presented in **Chapter 5**.
- **Chapter 3 (Applying Trustworthy Assurance)** presents findings from a research sub-project conducted with students and administrators from UK Universities. These engagement events explored the application of trustworthy assurance to the procurement of DMHTs for use in the higher education (HE) sector, as well as general attitudes and perceptions towards the use of data-driven technologies in higher education. A series of recommendations accompany our thematic analysis.
- **Chapter 4 (Co-Designing Trustworthy Assurance)** broadens the scope from the previous chapter to present research findings from a series of stakeholder engagement events carried out with regulators and policy-makers, developers, researchers, and users with lived experience of DMHTs. As with the previous chapter, a set of recommendations accompanies our thematic analysis.
- **Chapter 5 (Developing Trustworthy Assurance)** introduces, motivates, and explains two argument patterns that are intended to help project teams meet objectives for fair and explainable DMHTs. This chapter also connects the argument patterns to existing and relevant legislation and regulation (e.g. Equality Act 2010).

Examples

If you are new to the topics covered in this report, you can also find a set of illustrative examples of DMHTs available on [this page of our website](#).

About the Report

The following summarises what this report is and what it is not:

- ✓ An introduction to 'Trustworthy Assurance'—a framework and methodology for enabling a more trustworthy ecosystem of digital mental healthcare through the responsible and ethical design, development, and deployment of digital technologies.
- ✗ A comprehensive user guide for 'Trustworthy Assurance' or argument-based assurance—though links and further resources are provided.¹
- ✓ A summary of findings from research conducted on the application of trustworthy assurance to the procurement of DMHTs for use in the higher education (HE) sector.
- ✗ Findings from a sociological study or series of generalisable results from scientific experiments.
- ✓ A summary of findings from a series of more general stakeholder engagements, exploring the ethics of digital mental healthcare and attitudes towards trustworthy and untrustworthy technologies.
- ✗ A report with a strong international or multi-national focus. While we make reference to non-UK developments in this domain, our primary focus is on the UK. However, the methodology we present and many of the findings we discuss have value beyond the UK.
- ✓ An explanation and discussion of two argument patterns exploring the goals of fairness and explainability in the design, development, and deployment of DMHTs.
- ✗ A critical examination of argument-based assurance.²
- ✓ A series of recommendations, targeted at different stakeholders, for how to enable a more responsible and trustworthy ecosystem of digital mental healthcare.
- ✗ A review of the current legislative or regulatory publications that are relevant to digital mental healthcare.

Who is this report for?

This report is primarily targeted at the following groups:



Policy-makers and Regulators

Policy-makers and regulators will find the recommendations and guidance we set out of specific interest, and will find value in the methodology that we set out in [Chapter 2](#) because it is framed in procedural terms, and with links to process-based forms of governance. We also link our two argument patterns, which are presented in [Chapter 5](#), to specific legislative and regulatory developments in healthcare.



Senior Decision-Makers

Senior Decision-Makers, like policy-makers and regulators, will likely benefit from our methodology of Trustworthy Assurance. Specifically, from exploring its procedural underpinnings that are discussed in the section on a typical ML or AI lifecycle ([see Chapter 2](#)).



Developers and Product Managers

Although our framework and methodology are not set out using formal syntax and schemas for argument-based assurance, Trustworthy Assurance is, nevertheless, primarily aimed at developers and product managers. For instance, one of its key values is as a reflective and deliberative aid for demonstrating how ethical principles and decisions have been undertaken and establish through a project's lifecycle, with easy means for justifying the relevant claims by linking them to evidence. Therefore, developers and product managers are a key stakeholder group that we have targeted in this report and research project.



Researchers

Researchers may find less practical value in our framework and methodology than the above groups. However, our report highlights and emphasises significant knowledge gaps and research opportunities for improving our collective understanding about the individual and social impacts of DMHTs. Therefore, our report can also be seen as a call for further research into specific areas, including the evaluation and validation of the Trustworthy Assurance framework and methodology.

Users of DMHTs may also find value in the report, but it has not been produced with members of the public as primary target audience. In general, the responsibility for utilising the methodology and implementing and acting upon the recommendations is for the groups above; they are not the responsibility of the user!

The header features a dark blue background with various geometric shapes in light blue, teal, pink, and white. On the left, the numbers '01' are displayed in a large, white, sans-serif font. The rest of the header is filled with abstract shapes including circles, squares, and concentric circles in various colors and sizes.

01

Introduction

Laying the Foundations

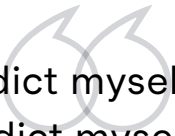
The Current (Socioeconomic) Landscape of Digital Mental Health

- › What is digital mental health technology?
- › Why and how is digital mental healthcare being used?

A Culture of Distrust

About the Project

- › SAFE-D Principles



Do I contradict myself? Very well
then I contradict myself, (I am large,
I contain multitudes.)

– Walt Whitman, *Song of Myself*

Whitman’s ode to self-knowledge and understanding contains many poetically-phrased truths. However, the one expressed in the above line is an understatement. If we were to identify and rank the most complex phenomena in the universe, our large and multitudinous minds would sit somewhere near the top of the list!

Even the most stubborn among us must acknowledge that part of this complexity stems from a capacity for our minds to operate as a network of often contradictory beliefs, attitudes, and opinions—a network that exists within and among a larger social network of similarly fallible individuals. It would be understandable, therefore, given this reflection, if we came to the conclusion that our minds were never supposed to be understood fully and we just accepted, as Whitman did, that our mental lives are fundamentally contradictory and diverse, and sometimes none the worse for it.

For many people, a prescription of stoic acceptance in the face of overwhelming complexity would be welcomed. But for others, their minds are not just built on top of permissible and tolerable contradictions, they also operate in a manner that prevents them from living a fully self-determined and flourishing life.

In the last decade or so, a wide range of digital and data-driven technologies have emerged that promise to improve both our knowledge and understanding of our complex minds and its capabilities, as well as enhance our overall well-being. This fact is unsurprising. Our species has used technology to learn about and restructure both our external and internal worlds for hundreds of thousands of years. And during our time on this planet, technology has both enhanced and diminished our knowledge, understanding, and individual and social welfare. So, why has so much attention been paid in recent years to a recurring cycle of technological innovation?

About this chapter

This introductory chapter provides contextual information for the report. However, it can be treated as an optional chapter for those readers who only want to engage with the trustworthy assurance methodology ([Chapter 2](#)), findings and analysis from our workshops ([Chapters 3 and 4](#)), or positive proposals ([Chapter 5](#)).



LAYING THE FOUNDATIONS

Towards the start of the 21st Century, a convergence of several social and technical factors gave rise, first, to an interest in the big data revolution, and, second, to a renewed interest in Machine Learning (ML) and Artificial Intelligence (AI). Let's look at each of these briefly, as they help establish important and explanatory context for this report.

The big data revolution occurred as a result of increased and widespread use of Internet of Things (IoT) or mobile devices (i.e. the sources of data); availability of affordable cloud computing infrastructure (i.e. for extracting, loading, and transforming the data); and ongoing development of open-source frameworks and software libraries for more efficient and distributed data storage and analysis (e.g. Apache Hadoop, Python), among other factors.

To help differentiate big data from ordinary data collection, analysis, and use, many have pointed to the five V's of big data:

VOLUME: — the <i>amount</i> of data being extracted	VARIETY: — the <i>types</i> of data being extracted	VELOCITY: — the <i>speed</i> at which data is extracted	VALUE: — the <i>socioeco- nomic benefit</i> of data	VERACITY: — the <i>accuracy</i> of data
--	--	--	---	---

In the context of digital mental healthcare, all of these are noteworthy, but three stand out against the backdrop of this report's opening remarks regarding the complex phenomena of interest (our minds):

-
- 1** How **accurate** are the data we are now collecting, analysing, and using?

 - 2** Given the **variety** of minds that populate this planet, how representative are the types of data?

 - 3** How much of a gap is there between the socioeconomic **value** of data and the value to the individual who is represented by the data? Or, to put it more bluntly, who benefits from the data?

We (the authors) have heard and discussed many variants of these questions over the course of this project. For example, concerns about accuracy and variety are deeply connected to considerations around the regulatory assessment of clinical efficacy and safety for novel data-driven medical devices.¹ But more than this, accuracy and variety also underpin broader ethical concerns about existing barriers to enabling a fairer and more accessible healthcare system (e.g. the digital divide that systematically excludes certain people and groups from benefits associated with digital technologies). However, although data are important, addressing these questions is just one part of the puzzle.

Turning to the technologies themselves—another significant piece—we can similarly identify several explanatory factors behind the recent surge of interest in machine learning algorithms (ML) and artificial intelligence (AI). Developments in this domain build on top of the aforementioned factors behind the big data revolution—ML and AI are, after all, sometimes referred to as ‘data-driven technologies’:

- › Theoretical advances in machine learning for robotics and intelligent software agents (e.g. DeepMind’s Alpha Go)
- › Improved application of deep neural networks to well-defined tasks such as medical imaging or speech detection
- › Hardware improvements in specialised computer processor architectures to allow for more efficient and effective edge computing

All of these developments are important, but again there are three aspects that stand out as significant:

1 The ability for ML/AI systems to operate **autonomously**

2 The ability for ML/AI systems to **learn** from their environments

3 The ability for ML/AI systems to **adapt to** and **affect** their environments

As we will see throughout this report, these features of ML algorithms and AI systems create possible risks and benefits to the realisation of ethical goals associated with digital mental healthcare. For instance, the ability to respect a patient’s right to autonomous decision-making. Furthermore, these issues intersect with the issues raised by the previous three questions pertaining to data (e.g. the ability to operate autonomously in complex environments with insufficiently accurate data).

These topics already paint a very complex picture, but there is also a further level of complexity involved with understanding the dynamic feedback loops that emerge in mental healthcare

when autonomous and adaptive systems are used to complement existing therapeutic interventions, many of which are already poorly understood (e.g. Selective Serotonin Reuptake Inhibitors). This complexity can cause issues for our existing research, development, and regulatory frameworks, such as when performing clinical trials (e.g. how should we control for the effects of adaptive and personalised technologies?).

Collectively, these six points about Big Data and ML/AI help to establish the background and context for this report, and also help us gain some conceptual clarity when attempting to address the uncertainty around trustworthy DMHTs. Some of this uncertainty stems from the technologies themselves (as noted above). But other key aspects of this uncertainty arise because a) the concept 'trustworthy digital mental health technology' is a poorly defined term² that captures a vast and heterogeneous class of tools and services, and b) our relationships to and interactions with the technologies are also varied. It is not just the technologies that are complex after all. As eloquently captured by Whitman at the start, we—the individual members of the class, 'humanity'—are large and contain multitudes.

These many layers of complexity coalesce into a particularly thorny problem. If we cannot trust the technologies themselves, we will not be able to trust the information gained from them about our own minds. But if we do not understand our minds, we may be unable to fully determine and address the cause of our trust or distrust. And, as we have just noted, understanding the social environment is also a vital part of addressing this fundamentally sociotechnical problem. As you can probably guess by now, adding another element to our picture is not likely to reduce its complexity.

THE CURRENT (SOCIOECONOMIC) LANDSCAPE OF DIGITAL MENTAL HEALTHCARE

To explore and understand the socioeconomic landscape of digital mental healthcare, let us start by addressing two outstanding questions:

1 What is meant by 'digital mental healthcare technology'?

2 Why and how are 'digital mental healthcare technologies' being used?

What is digital mental health technology?

This is not an easy question to answer because of the multifaceted ways that the term 'digital mental health technology' could be employed.

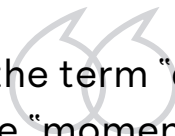
We can attempt to answer the question by drawing distinctions between, say, the use of digital technologies within formal healthcare settings (e.g. NHS) and those outside. However, as we have explored in previous research³, this boundary is vaguely drawn at best, and unhelpful at worst, when it comes to understanding DMHTs and the ethical issues surrounding their design and use.

One reason for this is that outside of formal healthcare systems, DMHTs have been employed in social domains and contexts as diverse as financial services, education (e.g. schools and universities), and employment. And, furthermore, DMHTs have been used within such domains for myriad purposes ranging from vulnerability assessment through to proactive intervention.³ There is also the use of DMHTs by social media platforms and charities to consider, which tends to cross many of these boundaries, especially those between work and home life, making it difficult to draw useful distinctions unless a narrow focus is defined in advance (e.g. studying the use of NLP used for risk assessment of adolescents on social media).¹³

What about focusing on the technologies themselves? Again, this is not an easy feat. Some organisations, such as the Nuffield Council on Bioethics have focused on emerging technologies to narrow their scope.⁴ In doing so, they have identified specific ethical challenges associated with the following class of technologies:

- › Smartphone apps and chatbots
- › Predictive analytics (e.g. based on digital phenotyping)
- › Consumer neurotechnology (e.g. portable electroencephalography devices)
- › Immersive technology (e.g. virtual reality)

But none of these categories are suitable for building a definition of DMHTs writ large. For instance, let's take a look at so-called "digital phenotyping"—an increasingly popular area of digital mental health—, defined as follows:



We also introduce the term "digital phenotyping" to refer to the "moment-by-moment quantification of the individual-level human phenotype in-situ using data from smartphones and other personal digital devices."⁵

Putting aside the previously raised data challenges associated with the "quantification of the individual-level human phenotype" (e.g. how to construct accurate scales that apply to all varied people while retaining value), there is also wide variation between 'smartphones' and 'other personal digital devices', such as wearables. For instance, some smartphone apps may use advanced forms of machine learning algorithms or AI to infer novel attributes regarding a user's mental health. And, others may offer nothing more than a simple interface and database for users to record how they are feeling at a particular time or on a particular day. Furthermore, some wearables may store and process sensitive information locally on a user's device, whereas others may store data on the cloud and share health-related information with a vast number of organisations across jurisdictions with varying levels of data protection.

What about if we move to a lower level of abstraction, such as the algorithmic technique used by the digital technology? Here too we would find difficulties with delineating the meaning of the term 'digital mental health technology'. For example, the use of unsupervised machine learning by trusted clinical researchers may be justified if used responsibly as a form of exploratory research or hypothesis generation. But, if a complex version of the technique were deployed by a local council to determine how best to spend limited resources (e.g. resulting in clusters that were not clearly interpretable by humans), the potential lack of transparency could undermine efforts to remain accountable to their residents.

Regardless of the level of abstraction we adopt, there will always be some difficulty with clearly defining this nebulous term. Therefore, while it may seem unsatisfying to a reader who wishes to know precisely what the term 'digital mental health technology' comprises, for our present purposes the following (loose and permissive) definition shall suffice:



The term 'digital mental health technology' refers to any digital technology that has been designed, developed, and deployed with the goal of improving or otherwise impacting some mental health outcome for an individual or group of people.

In particular, this report will pay close attention to those technologies that are data-driven and/or use some form of machine learning or AI, given the considerations outlined in the opening sections.

We acknowledge that many will find this definition too permissive, but this report is not concerned with developing a robust philosophical definition or a taxonomy that can be used to delineate the precise nature of DMHTs. Rather, it is focused on the defence of a methodology to help make DMHTs more trustworthy and ethical. Therefore, the broader the class that can be drawn the better it will be for our goals, because more technologies will fall within its scope⁶. And, insofar as there are legal considerations that demand precise definitions, these issues will be addressed along different lines (e.g. operationalising standards of assessment for equitable treatment).

Let us now turn to the second question.

Examples

A set of illustrative examples of DMHTs are included in the online version of our report. You can access them [here](#).



Why and how is digital mental healthcare being used?

The following statistics offer a partial and fragmented perspective to help frame this question, focusing on the UK specifically:

- › Over 60% of children and young people with diagnosed mental health conditions do not receive NHS care.⁷
- › Rates of probable mental health disorders in children and young people (aged 6 to 16 years) have risen from 11.6% in 2017 to 17.4% in 2021.⁸
- › Approximately two-thirds of people who die by suicide are not in contact with NHS mental health services.⁹
- › In the first 3 months of 2021, 1 in 5 adults in Britain experienced some form of depression (over double the pre-pandemic figures).¹⁰

- › During the pandemic both males and females saw an increase in anxiety and a reduction in 'life satisfaction'—a subjective measure of well-being that asks individuals to evaluate their life as a whole, rather than time-specific emotions. However, females experienced lower life satisfaction and happiness than males.¹⁰

When we consider these figures, combined with a reflection of the impact wrought by the COVID-19 pandemic on an already over-burdened mental health sector, we can begin to understand why many organisations across the public, private, and third sectors are deploying digital technologies to augment and complement their services, and why many users in turn have engaged.

But if we are to implement digital technologies in an ethical and trustworthy manner, there are several considerations that need to be addressed.

The first two relate to choice and access, as outlined in a briefing note from the Nuffield Council on Bioethics⁴ (emphasis ours):

1 “Many people affected by mental health problems do not have access to or are reluctant to use mental healthcare technologies. If these are to become widely adopted in the future, there should be choice about using them.”

2 “Technology solutions should not divert resources from other important forms of mental healthcare and support and should be used as an addition to what is already available, rather than a replacement.”

The latter conclusion is echoed here because it is a theme that emerged frequently in our own project among diverse stakeholder groups. That is, DMHTs should augment and support, but never replace human decision-making or human-centred services. And, the former conclusion is also important because it captures something salient about trust.

For some potential users, such as elderly patients, a lack of access can be due to their needs not being sufficiently considered when designing the service or technological interface.¹¹ This form of inaccessibility is sometimes overlooked due to an emphasis on other economic barriers (e.g. digital poverty). However, even users that a) have access to the services (in both senses of the term 'access), and b) potentially benefit from use of the respective technology, may still have legitimate reasons for not wishing to use the service due to a distrust (or, “reluctance” to use the same term from the above quotes) in the service or the organisation responsible for designing, developing, and deploying it. In some cases, this distrust arises due to legitimate concerns about violations of data privacy or mishandling of sensitive information by commercial organisations.

However, the unethical behaviour and transgressions of law by commercial organisations such as Facebook¹² can have a wider impact beyond their own disastrous public relations. They can also contribute to a growing culture of distrust in the ecosystem more broadly, affecting the public and third sectors, as members of the public may be unable to separate the differing ethical, social, or legal norms that govern each sector or domain.¹³ This is understandable from the perspective of the user, as the norms that regulate and govern the public, private, and third sectors are complex and deeply interwoven. But it is still characteristic of an unethical and irresponsible approach to research and innovation, and one that is unlikely to build trust.

A CULTURE OF DISTRUST

In the context of the law, it is well known that states and public sector organisations are beholden to wide-ranging legal duties, both positive and negative, such as those set out in human rights law or in national legislation (e.g. the public sector equality duty created by the UK's Equality Act 2010).

Commercial organisations are not obligated to observe all of the same principles or rules as public sector organisations, but are nevertheless required to comply with myriad information governance standards, legislation designed to protect environmental sustainability and public health, and a whole host of other corporate or fiduciary duties.¹⁴

Third sector organisations, such as charities or volunteer groups, may have less restrictive legislation governing their conduct, but are still expected to adhere to necessary transparency and accountability standards over matters such as the organisation and incorporation of managing trusts.

Such legal requirements create an interlocking foundation upon which public perceptions and attitudes towards trust can be based, but are often difficult to separate and pick apart. And, even where one is able to do so, legal requirements typically set only the minimal standards expected of organisations. To put it simply, and sidestep a vast amount of important jurisprudence, just because something is legal does not guarantee it is ethical or socially acceptable.

On top of the norms that fall within the scope of the law, modern institutions and organisations are also expected to observe and comply with an expansive and shifting set of ethical and social norms. For example, while underpinned by legal mechanisms, matters of social justice and fairness go beyond the legal requirements to ensure non-discrimination (e.g. poverty, a risk factor associated with worse mental health outcomes, is not a protected characteristic¹⁵ as set out in the Equality Act 2010).¹⁶ Moreover, legal texts often leave wide scope for actions that may be sufficient to discharge duties corresponding to individual rights¹⁷, but are seen by many as, at best, failing to observe the spirit of the law, and at worst, morally impermissible (e.g. privacy policies).

A particularly well known illustration of this problem is the EU's General Data Protection Regulation (GDPR) and ePrivacy Directive. The GDPR (and directive) resulted in widespread changes to the operation of cookies, including a requirement to receive users' consent before any cookies were used, except those strictly necessary. However, as almost everyone will know from first-hand experience, the manner in which some organisations secure consent can range from the entirely user-friendly, to the intentionally frustrating use of dark patterns¹⁸ or hours long process of flipping hundreds of opt-out toggle buttons. Here, the expectations that society have regarding what is both legally and morally permissible clearly differ substantially from what is desirable from the perspective of the organisation and the law.

But these expectations also differ depending on whether the organisation is part of the public, private, or third sector, and what role they play within each sector. And, furthermore, expectations are not equally shared across a vast and homogenous “public”. Quite the opposite in fact.

Consider, for example, the range of attitudes that members of the public may have towards a private company extracting economic value from their data collection activities. Depending on key details about the informational content of the data, attitudes could range from shareholder praise for savvy corporate governance, through to begrudging toleration by consumers, and up to the vehement and vocal criticism by privacy activists or employee campaign groups.

And to add one final layer of complexity, to really drive home Whitman’s point from the start of this chapter, the scope and distribution of this variation may increase as we expand our field of consideration to the public and third sectors. Now, the same data extraction could be seen as deeply unethical or impermissible by those who were in favour of it originally. In terms of underlying values, therefore, *pluralism* and *variation* should be expected when considering the attitudes of the publics (reiterating the emphasis on the plural).

By now, you may be feeling as though stoic acceptance in the face of overwhelming complexity is inevitable. What other options are there? You may be thinking, for instance, that there are simply too many factors for any one person or organisation to consider when researching, developing, or regulating DMHTs.

However, this report (and our project more generally) aims to challenge this attitude while fully acknowledging the overwhelming complexity involved. Our approach and methodology in many respects embodies the principle that light is the best disinfectant. We can frame our approach as a set of recommendations that build on the two earlier conclusions from the Nuffield Council on Bioethics (see next page).

Original Conclusions from Nuffield Council on Bioethics

- 1** Many people affected by mental health problems do not have access to or are reluctant to use mental healthcare technologies. If these are to become widely adopted in the future, there should be choice about using them.
- 2** Technology solutions should not divert resources from other important forms of mental healthcare and support and should be used as an addition to what is already available, rather than a replacement.

Additional Recommendations

- 3** Organisations that choose to use DMHTs should consider broader ethical goals, in addition to traditional goals such as 'safety' and 'efficacy', to help create a more ethical, responsible, and trustworthy ecosystem of digital mental healthcare.¹⁹
- 4** Organisations should provide transparent and evidence-based assurance about how these ethical goals have been operationalised and secured during the design, development, and deployment of DMHTs.

With these points in mind, we can now turn to the project itself and introduce the notion of trustworthy assurance—a methodology that can help address all of the above recommendations (and more to come).


ABOUT THE PROJECT

Assurance is a process of establishing trust. Whether we trust someone or some object depends, in part, on the evidence we have to help us evaluate whether there are good grounds for placing trust. In other words, what is the evidence of their trustworthiness?

When it comes to trust, we do not expect the same level of evidence when assessing the trustworthiness of different people, objects, or systems. A trustworthy doctor, for example, is not assessed by the same standards as a trustworthy friend. And, similarly, the trustworthiness of an AI chatbot used in customer services is not (and ought not) be evaluated by the same measures as an AI chatbot used to support people with their mental health.

In short, when we speak of 'trustworthy assurance' we are creating room for a wide variety of associated goals and standards, to accommodate the complexity alluded to in the previous sections. These can, of course include goals and standards related to 'safety' or 'clinical efficacy', which carry their own ethical significance. However, for present purposes we are primarily interested in those goals that are directly framed in terms of ethical principles (e.g. fairness).

Our project focused directly on a methodology for making the assessment, communication, and realisation of these goals more robust and transparent. The methodology is known as 'argument-based assurance' (ABA) and we can define this methodology 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.²²

We offer a simplified introduction to ABA in the following chapter. But various types of ABA are already widely used in safety critical domains, and have also been used in the context of healthcare²⁰. Typically, the purpose of ABA is to assess and communicate the safety of a system within a particular environment. Our project was concerned with the question of whether a revised and extended version of the methodology could be used for a broader set of ethical goals, such as fairness or explainability (see [Chapter 5](#)).

There is an immediate question that ought to be addressed here:



How should ethical goals be determined and operationalised in the context of the design, development, and deployment of DMHTs?

Our approach in this project to determining and operationalising the relevant ethical goals was participatory in nature, and was driven by three primary objectives:

-
- 1** To explore whether and how the methodology of ABA could be extended to address ethical issues in the context of digital mental healthcare.

 - 2** To evaluate how an extension of the methodology could support stakeholder co-design and engagement, in order to build a more trustworthy and responsible ecosystem of digital mental healthcare.

 - 3** To lay the theoretical and practical foundations for scaling the trustworthy assurance methodology to new domains, while integrating wider regulatory guidance (e.g. technical standards).

To realise these objectives, several workshops were organised and run over the course of the project with a diverse set of participants. Broadly, we categorised these stakeholder groups as follows:

- University students
- University administrators
- Policy-makers and regulators in healthcare
- Developers of DMHTs
- Researchers working in disciplines adjacent to digital mental healthcare
- Users with lived experience of DMHTs

Workshops and interviews were held with representatives from each of these stakeholder groups, where tailored activities were run to both understand their attitudes towards DMHTs, but also to a) help us evaluate methodological questions related to trustworthy assurance and b) identify which ethical values and principles matter most to them in the present context.

Chapters 3 and 4 present our findings, analysis, and recommendations from the engagements. Here, we shall just speak to the procedural matter of operationalising ethical principles through processes of stakeholder participation and engagement.

SAFE-D Principles

In previous work, we have defended an ethical framework for evaluating the harms and benefits of data-driven technologies, which has already been revised, tested, and validated with a wide-variety of stakeholders.²¹

We refer to this framework as the SAFE-D framework, because it establishes five principles that form the acronym SAFE-D (or, 'safety', which is another important component of trustworthy AI):

- ➔ **Sustainability**
- ➔ **Accountability**
- ➔ **Fairness**
- ➔ **Explainability**
- ➔ **Data** (Quality, Integrity, Protection and Privacy)

Each of the SAFE-D principles has a subset of core attributes that help to specify and operationalise the principles throughout a project's lifecycle using a series of processes and activities (see [next chapter](#) for full details).

In other words, while the principles themselves act as starting points for context-specific reflection and deliberation with affected stakeholders, it is the core attributes that serve as practical guardrails throughout a project's lifecycle. For instance, the principle of 'explainability', which emphasises core attributes such as transparency, interpretability, and accessibility of an automated system, has a particular ethical significance when utilised in a domain such as digital mental healthcare. That is, ensuring digital mental healthcare technologies and services are explainable is a key part of respecting a patient's right to informed and autonomous decision-making. This right cannot be upheld and respected without ensuring sufficiently transparent, interpretable, and accessible forms of information about how a digital technology operates (e.g. how an algorithmic system reaches a decision). How organisations achieve this goal is something this project and report addresses directly.

While the SAFE-D principles have been designed and refined over multiple years (in a domain-general context) their relevance in digital mental healthcare had, hitherto, not been evaluated. Therefore, part of this project involved the following:

-
- 1 Understanding which, if any, of the SAFE-D principles were significant to different groups of stakeholders, and whether specific core attributes could be identified and developed in conjunction with stakeholders.

 - 2 Identifying if there were any gaps or omissions in the SAFE-D framework.

 - 3 Determining whether any of the revised and domain-specific principles or attributes could serve as top-level goals or property claims in trustworthy assurance cases (see [Chapter 2](#)).

Our findings and analysis that address these specific research questions comprise the majority of [Chapters 3](#) and [4](#). Among other findings and recommendations, these sections show there is strong evidence to suggest that the methodology of trustworthy assurance will lead to positive impacts in digital mental healthcare, and help foster a more responsible ecosystem of research and innovation.

Before we discuss these findings and analysis though, it is necessary to introduce and explain the methodology of trustworthy assurance, which is the topic of the next chapter.

The header features a light blue background with various geometric shapes: a large dark blue '02' on the left, several smaller squares and circles in white and blue, and a series of concentric circles in the top right corner.

02

Presenting Trustworthy Assurance—A Framework and Methodology

Designing, Developing, and Deploying Trustworthy Digital Mental Health Technologies

› Reflective and Anticipatory Deliberation

What is Trustworthy Assurance?

- › Argument
- › Procedure
- › Standards

Argument Patterns

- › Claims as Reasons
- › What are argument patterns?
- › Generalisable Patterns

Chapter Overview

This section introduces a framework and methodology for enabling a more trustworthy ecosystem of digital mental healthcare through the responsible and ethical design, development, and deployment of digital technologies. The section also serves as an introduction for the analysis and recommendations for the following sections.

First, we introduce a model of a typical research or innovation lifecycle for a data science or AI project that includes activities of project design, model development, and system deployment.

Second, we discuss the methodology of trustworthy assurance that is at the centre of our project. We provide a simple overview of the relevant procedures, focusing on the structure, elements, and purpose of an assurance case.

Finally, we look at argument patterns: reusable templates that can be developed to ensure a more consistent approach to trustworthy design, development, and deployment of digital mental healthcare.



DESIGNING, DEVELOPING, AND DEPLOYING TRUSTWORTHY DIGITAL MENTAL HEALTH TECHNOLOGIES

Designing, developing, and deploying an AI system is not a one-person task!¹ The stages and activities that comprise a typical AI project lifecycle involve a wide-ranging set of skills and capabilities. These skills are encapsulated within a variety of roles, including 'project commissioner', 'product manager', 'data protection officer', 'data scientist', 'system architect' and 'software engineer'. And, these roles are interwoven such that they create an irreducible and collective responsibility that spans the entire project lifecycle, and may span multiple teams and organisations.

Figure 2.1 presents a simplified model of a typical research or innovation lifecycle for a data science or AI project lifecycle, to help gain a purchase on these interweaving roles, skills, and responsibilities².

The model represents three over-arching stages of (project) design, (model) development, and (system) deployment. For each stage, there are corresponding activities, detailed in **Table 2.1**. The project lifecycle is depicted as a circular process to highlight the fact that responsibility is ongoing and does not end once a system has been implemented or put into deployment. Rather, responsible (and trustworthy) approaches to research and innovation require consideration of how a technological system may need to be monitored and updated once in production, and removed and replaced once it reaches the end of its lifecycle.

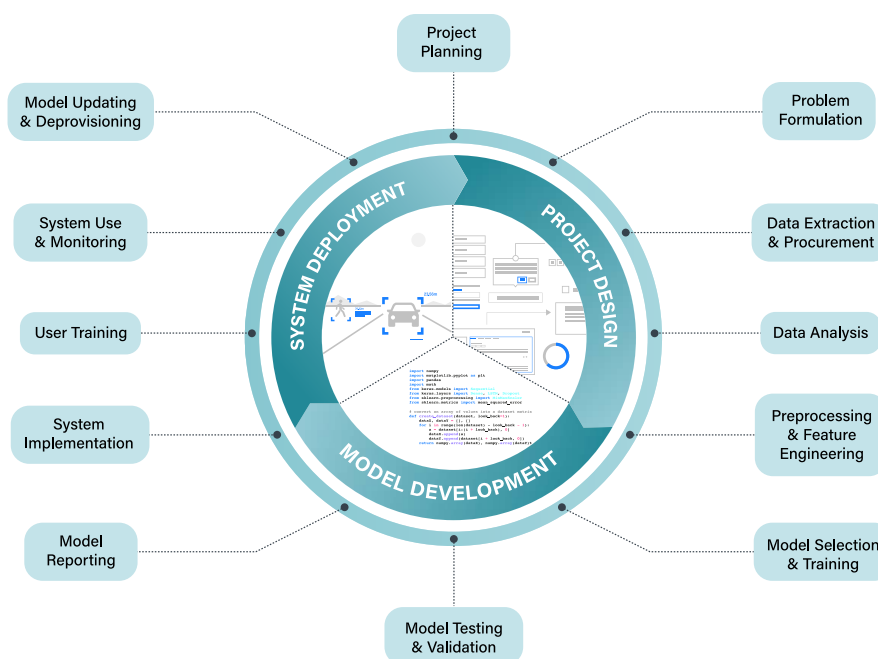


Figure 2.1: A model of a typical project lifecycle for a data-driven technology, detailing the overarching stages of design, development, and deployment and their constitutive activities (reprinted from [Burr and Leslie, 2022](#))

Table 2.1

Project Lifecycle Activities	
STAGE	DESCRIPTION
Project Planning	Preliminary activities designed to help scope out the aims, objectives, and processes involved with the project, including potential risks and benefits.
Problem Formulation	The formulation of a clear statement about the over-arching problem the system or project addresses (e.g., a research statement or system specification) and a lower-level description of the computational procedure that instantiates it.
Data Extraction or Procurement	The design of an experimental method or decisions about data gathering and collection, based on the planning and problem formulation from the previous steps.
Data Analysis	Stages of exploratory and confirmatory data analysis designed to help researchers or developers identify relevant associations between input variables and target variables.
Preprocessing and Feature Engineering	A process of cleaning, normalising, and refactoring data into the features that will be used in model training and testing, as well as the features that may be used in the final system.
Model Selection and Training	The selection of a particular algorithm (or multiple algorithms) for training the model.
Model Testing and Validation	Testing the model against a variety of metrics, which may include those that assess how accurate a model is for different sub-groups of a population. This is important where issues of fairness or equality may arise.
Model Documentation	A process of documenting both the formal and non-formal properties of both the model and the processes by which it was developed (e.g., source of data, algorithms used, evaluation metrics).

STAGE	DESCRIPTION
System Implementation	The process of implementing the technological system into its intended environment or target domain to enable and structure interaction with the underlying model(s) (e.g. a recommender system that suggests possible treatment options for patients based on input data).
User Training	Training for those individuals or groups who are either required to operate a data-driven system (perhaps in a safety critical context) or who are likely to use the system (e.g. healthcare professionals, medical researchers).
System Use and Monitoring	Ongoing monitoring and feedback from the system, either automated or probed, to ensure that issues such as model drift have not affected performance or resulted in harms to individuals or groups.
Model Updating or Deprovisioning	An algorithmic model that adapts its behaviour over time or context may require updating. ³ Where no further updating can be carried out, and this results in a system being removed from production (i.e. deprovisioned), a new system may be required. This restarts the project lifecycle.

To see how this model can help us understand the interwoven nature of responsibility, consider the following example. An organisation wants to implement a speech recognition algorithm within a service they are developing for online counselling. However, there is no one in the organisation with the relevant expertise to collect data and train a model from scratch. Therefore, they choose to procure a pre-trained model from another company. This means that a significant portion of the project lifecycle—from **Data Extraction or Procurement** to **Model Documentation**—will have been carried out by a separate organisation.

Although the specifics of the relationship between the two organisations will complicate forms of responsibility, such as legal duties or obligations, this need not concern us here. Instead, we can focus on how the initial organisation who has chosen to undertake the project (e.g. the product owner or commissioner) can use the project lifecycle model to a) identify and analyse their own responsibility and how it intersects with the responsibilities of others, and b) how this necessitates a process of trustworthy communication and assurance.

Reflective and Anticipatory Deliberation

At the start of a project, while activities such as planning and initial evaluation of feasibility are being conducted, the project lifecycle model can be used to structure *reflective* and *anticipatory* processes of deliberation among the project team. For instance, the team could use the model to identify and evaluate potential actions and decisions that are likely to emerge during specific activities, such as which data types may be required and whether stakeholders or users will consent to these data being collected and analysed (a reflective and anticipatory exercise). As this example suggests, the project team may carry out the preliminary deliberation, but additional stakeholders will need to be engaged to thoroughly evaluate the ethical, legal, and social permissibility and acceptability of the project.

Consider another example. A team of developers working for a commercial organisation have identified a risk associated with an AI system they have developed, which they claim is able to detect emotions. They have been approached by a healthcare provider who wish to procure and implement their system into a video consultation service to help their counsellors better understand the emotional and behavioural responses of their patients during an initial assessment. However, the developers did not evaluate their model (during **Model Testing or Validation**) using a dataset that is representative of the patient population intended by the healthcare provider that has approached them (i.e. individuals that are likely to be suffering from a mental health issue). Therefore, the developers are unable to make any claims about the generalisability of their model to this new population. Moreover, neither the developers nor the healthcare provider have engaged the relevant stakeholder groups during **Project Planning** to determine if this would be an acceptable use for their system. As such, additional activities would need to be carried out to determine the full scope of the risks and possible harms that could arise from the use of this technology. This would likely require the procuring organisation (i.e. the healthcare provider) to set clear requirements for what forms of evidence would be required from the developers (e.g., at **Model Documentation**), and to determine clear boundaries and thresholds for whether the project should proceed.

As this example illustrates, the project lifecycle structure can help support forms of reflective and anticipatory deliberation that help instantiate a responsible ecosystem of research and innovation. And, it can also help identify points in the lifecycle where structured and transparent communication between teams and organisations may be crucial.

In addition, there is a further purpose for the project lifecycle model that will become clearer in the next sections: the identification of actions and decisions that generate forms of evidence that provide justificatory support for *trustworthy assurance*.

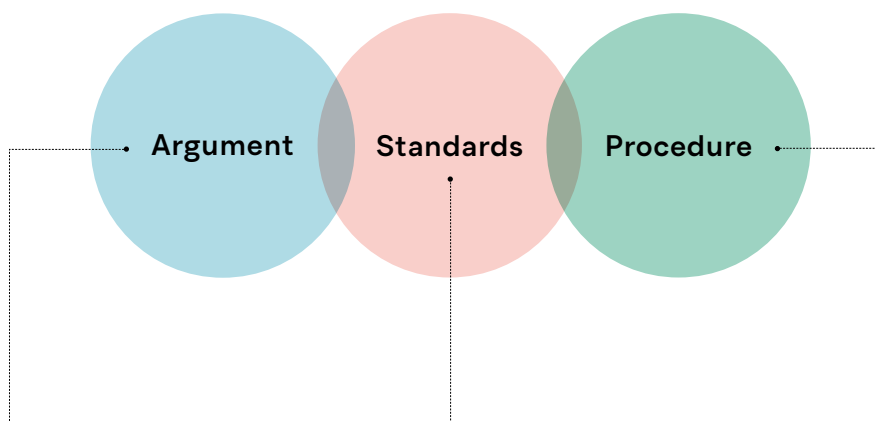
WHAT IS TRUSTWORTHY ASSURANCE?



Trustworthy assurance is a procedure for developing a structured argument, which provides reviewable (and contestable) assurance that a set of claims about the ethical properties of a data-driven technology are warranted given the available evidence.

This definition captures three important and interlocking components of trustworthy assurance:

- 1** A structured *argument* comprising linked claims and evidence that collectively justify a top-level goal
- 2** A *procedure* for developing an assurance case, which represents the argument either formally and/or visually
- 3** Agreed upon *standards* for reviewing and evaluating the argument



Generalisable structure to facilitate communication and best practices (i.e. argument-based assurance)

Evidential standards to ground generalisable claims

Transparent goal-directed process to provide accountability and build trust

Figure 2.2: A schematic showing the three interlocking components that support trustworthy assurance.

Argument


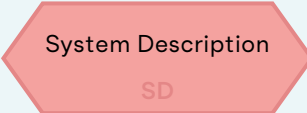
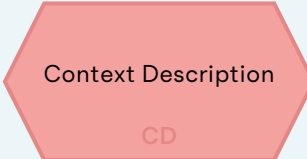
Trustworthy assurance is a form of argument-based assurance. It uses a structured type of documented argumentation, known as an assurance case, as the primary means for providing assurance that a goal has been obtained, based on the claims and evidence presented.

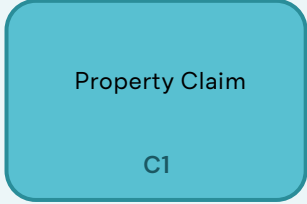
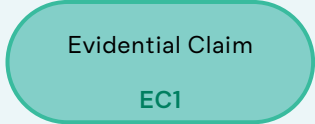

There are three basic elements of an argument:

- 1 A claim about the goal to be established (supported by descriptions of the system and the context in which the system is intended to operate)
- 2 A set of property claims about the project or system that collectively specify and operationalise the goal
- 3 A set of evidential claims that jointly establish the validity of the property claims

Box 2.1: Elements of a Trustworthy Assurance Case

The following table provides summary information about the typical elements of a trustworthy assurance case. NB: the colours do not mean anything. They are purely used as a visual aid to help differentiate the elements.

ELEMENT NAME	DESCRIPTION	ICON
Goal Claim	A claim about an ethical goal of the DMHT, which the assurance case attempts to justify has been established on the basis of the evidence and argument provided.	
System Description	A short description about the DMHT, including any central algorithmic techniques.	
Context Description	A short description about the intended context of use for the DMHT, including the users of the system (e.g. healthcare professionals).	

ELEMENT NAME	DESCRIPTION	ICON
Property Claim	A claim about how the ethical goal has been implemented or operationalised, which references a property of the system or project, e.g. an action that was undertaken during the model's development.	
Evidential Claim	A specific claim about some evidence, which serves to establish the validity of the higher level property claim.	
Evidential Artefact	A description of the evidence referred to by the above evidential claim, including a link to the relevant document where available.	

The following figure constitutes a simple (but incomplete) argument, showing the relationship between the three central elements.

Here, the goal that is being established relates to the project team's ambition to 'respect user privacy'. And, they argue that this is achieved by adherence to data minimisation principles—a claim about a property of how the system operates. Evidence of this adherence is also provided.

While useful as an illustration, this example is too simple to constitute a full-fledged assurance case because it reduces the concept of 'respect for user privacy' to a single principle (i.e., data minimisation). Although this claim may be relevant, on its own it is insufficient. We can, for example, consider an app that collects no personal data but still violates reasonable expectations of privacy by routinely notifying and disturbing users.

A more detailed procedure is required, therefore, to help project teams identify the set of property claims that a) specify and operationalise the top-level goal and b) collectively justify the goal.

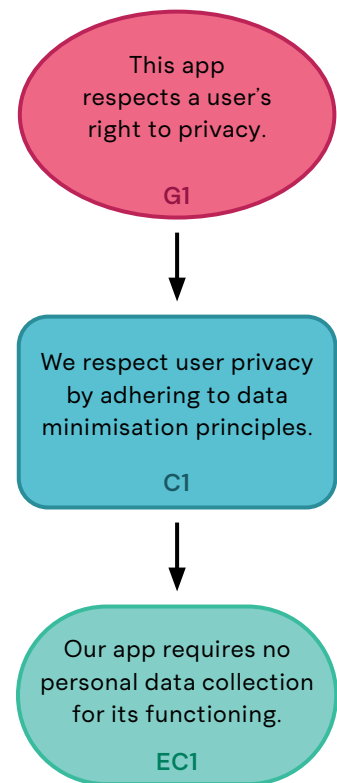


Figure 2.3: A portion of a simplified assurance case on respect for privacy

Box 2.2: Argument-Based Assurance Notation

Argument-based assurance (ABA) was defined in the previous chapter as follows:



“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.”

ABA is widely used in safety-critical domains or industries where manufacturing, development, and maintenance processes are required to comply with strict regulatory requirements and legislation, while also supporting industry-recognised best practices¹⁶. Because of these requirements, there are many formal standards that can be used to better govern the process of constructing an assurance case.

A popular standard is known as ‘Goal Structuring Notation’ (GSN)—originally developed in the 1990s at the University of York to assist the production, maintenance, and reuse of safety and assurance cases in safety critical industries such as traffic management and nuclear power¹⁷. There are many similarities between GSN’s assurance cases and Trustworthy Assurance, as the latter was directly inspired by the former. For example, as the name implies, GSN structures an assurance case towards a particular goal, and best practices associated with the standard prescribe methods for minimising possible harms are proportionate to the risks presented by the technology or system (e.g. minimisation of safety risks to levels that are as low as reasonably practicable). However, GSN also has additional elements (e.g. solutions and strategies) and relationships between elements that we do not include in the current presentation of Trustworthy Assurance⁴.

It should also be noted that our use of colours for the various elements should not be seen as signifying any meaning within a formal context. This choice was made solely for ease of comprehension for our stakeholders who were unfamiliar with the method. As we discuss in the [Conclusion](#), our future ambitions are to explore how GSN can be used to anchor Trustworthy Assurance in a more formal notation or schema. However, for this project we chose to sidestep many of the formal considerations that arise in the GSN standard (or any other formal standards⁵) due to the likely barriers to comprehension that existed for our stakeholders.

Further Resources

The following resources provide good overviews and clear introductions for the reader who is interested in further exploring argument-based assurance:

- > Kelly, T. (1998) Arguing Safety – A Systematic Approach to Managing Safety Cases (PhD Thesis). Available: <https://www-users.cs.york.ac.uk/~tpk/tpkthesis.pdf>
- > The Assurance Case Working Group. (2021). GSN Community Standard Version 3. Available: <https://scsc.uk/r141C:1?t=1>
- > Hawkins, R., Paterson, C., Picardi, C., Jia, Y., Calinescu, R., & Habli, I. (2021). Guidance on the Assurance of Machine Learning in Autonomous Systems (AMLAS). University of York. <https://www.york.ac.uk/media/assuring-autonomy/documents/AMLASv1.1.pdf>
- > Laher, S., Brackstone, C., Reis, S., Nguyen, A., White, S., & Habli, I. (2022). Review of the Assurance of Machine Learning for use in Autonomous Systems (AMLAS) Methodology for Application in Healthcare. 32. <https://arxiv.org/ftp/arxiv/papers/2209/2209.00421.pdf>
- > Sujan, M. A., Habli, I., Kelly, T. P., Pozzi, S., & Johnson, C. W. (2016). Should healthcare providers do safety cases? Lessons from a cross-industry review of safety case practices. *Safety Science*, 84, 181–189. <https://doi.org/10.1016/j.ssci.2015.12.021>

Procedure

The procedure advocated for trustworthy assurance is the anticipatory and deliberative exercise introduced above, which incorporates inclusive and accessible forms of *stakeholder engagement*.

By using the project lifecycle model as a scaffold for anticipatory reflection and stakeholder-informed deliberation, project teams are able to answer the following questions:

-
- 1 Which claims (that may emerge from participatory deliberation) are necessary and sufficient to specify and justify the top-level goal?
-
- 2 How do these claims relate to one another?
-
- 3 What evidence is required to demonstrate the validity of the claims being made?

Consider the following example. A company wishes to develop an assurance case that shows how their system, which uses a ML algorithm to predict whether users of an online betting platform are “problem gamblers”, can generate results that are explainable to their users. They decide this is an important ethical goal for an assurance case, because they want to be able to provide accessible forms of communication to any user that they contact on the basis of their algorithmically-generated prediction.

They start by formulating the following goal statement, which sets out an ambitious objective to achieve:



“An explanation of how our system predicts whether a user is a “problem gambler” can be provided to all users of our platform.”

Next, they consider which potential actions or decisions taken throughout the stages of the project lifecycle could be relevant to the specification and justification of this goal. For example, they flag that results from a series of planned workshops to be carried out with representative users during their **Project Planning** activities may be relevant.

Following the delivery of these workshops, it turns out that plain language explanations are preferable to detailed explanations of the algorithmic techniques, but that users were more trusting of these explanations when they knew they had been independently validated by a professional auditor. This result influences which machine learning algorithm the team go on to select during the **Model Selection and Training** activities, and which features they report on during **Model Documentation** for the independent audit.

Once the team have reflected on all the stages of the project lifecycle, and carried out the corresponding activities, they recognise that there are two broad sets of claims. One set of claims are about the design choices made during the project, which support accessible explanations for users. The second set are about the *interpretability* of the system, which are relevant for professional auditing and assessment of the system. Categorising the claims in this manner helps the company determine what evidence will be needed for each claim, and how best to structure the argument. At this point, most of the evidence has already been generated as a byproduct of the team’s work, so this stage is primarily a matter of collection, curation, and communication (through documenting an assurance case).

The following figure summarises the points made in the above example.

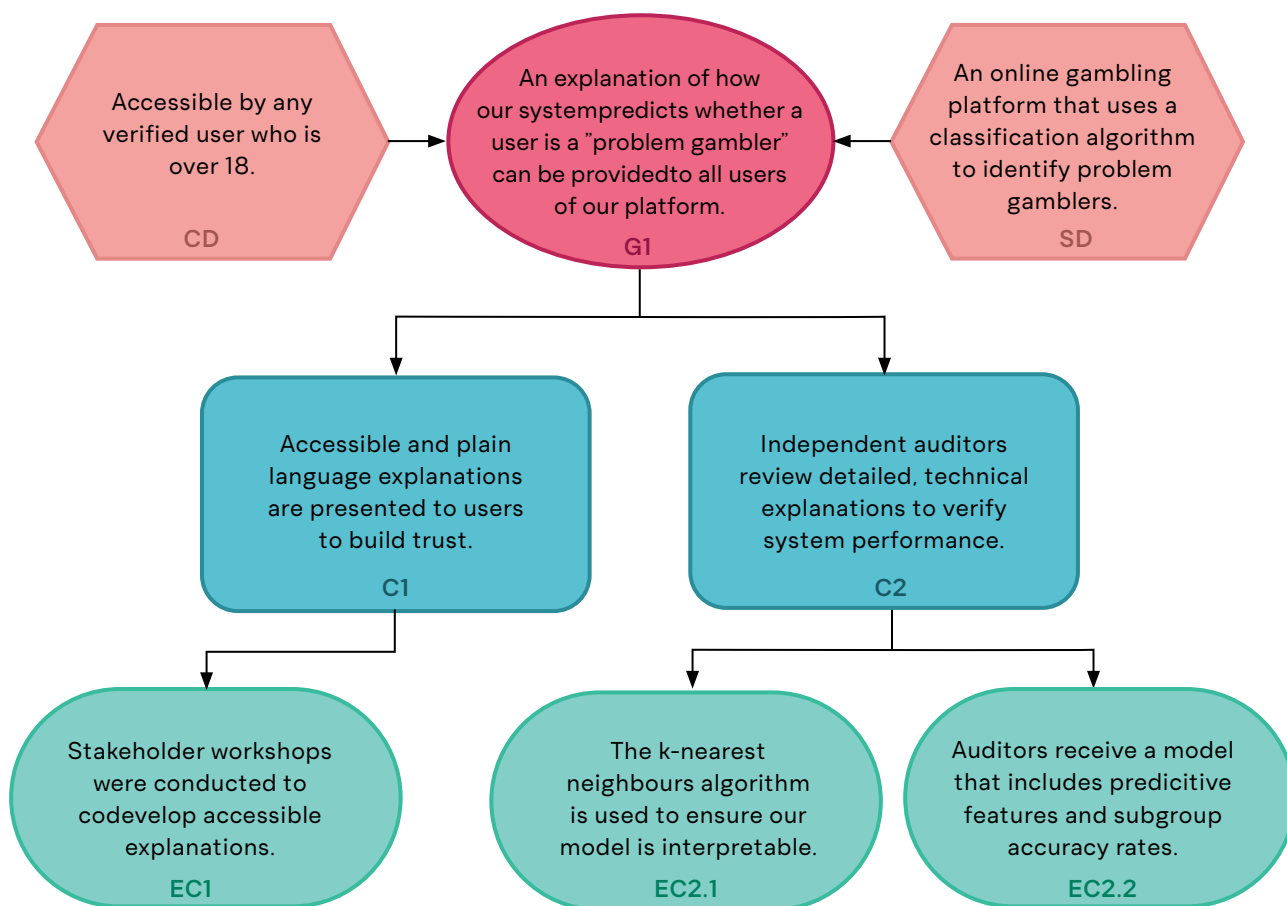


Figure 2.4: A portion of the assurance case for this hypothetical project.

The final component of trustworthy assurance relates to standards.

Standards

Standards support the development and refinement of best practices and codes of conduct. There are standards for measurement (e.g., universal scales), procedures (e.g., manufacturing), and assessments (e.g., risk and impact assessment), and much more. Here, we are interested in standards as they pertain to evidence and claims.

Evidential standards can refer to both the identification and evaluation of evidence.

Standards for identifying evidence are common in areas such as law where rules exist to determine what constitutes relevant, material, and admissible evidence.

Similarly, standards for evaluating the quality of evidence are well-established in domains such as scientific research, where various procedures or methods are held to produce reliable forms of evidence (e.g. peer review or randomised controlled trials), and in risk assessment and management (e.g. standardised guidelines on risk management for systems and software engineering).⁶

As distinct communities of practice develop and emerge within digital mental healthcare, we would expect standards and best practices to evolve to help with both the identification and evaluation of evidence. Subsequently, this would help support the development of trustworthy assurance in domains such as digital mental healthcare because specific types of claims or evidence would be recognised as more reliable forms of evidence-based assurance. It should be noted, however, that regulators and developers are not starting with a blank slate. There are many relevant standards that exist today, and new standards are emerging to support the procedure of constructing a trustworthy assurance case.⁷ We will consider some of these standards in [Chapter 3](#).

ARGUMENT PATTERNS

Claims as Reasons

Trustworthy assurance is a process of giving and justifying claims about choices made during the design, development, and deployment of DMHTs. These claims can be viewed as a series of reasons for why a particular decision was made.

To see why, let's assume that an organisation is in the process of procuring an AI-enabled chatbot to provide therapeutic support to service members returning from deployment.⁸ As this technology is new and relatively untested, the organisation has a series of questions for the developers:

-
- 1 "Why should we license this digital system instead of investing in traditional forms of talk therapy?"

 - 2 "Why have you chosen a female avatar as your virtual assistant?"

 - 3 "How did you measure and validate the clinical efficacy of the system for different subgroups to ensure that it is fair?"

In answering these three questions, the developers would be giving reasons (supported by evidence) for their actions—reasons that would need to be accepted by the procuring organisation to be relevant and justifiable. This perspective on claims emphasises one of their most vital roles as publicly contestable reasons. That is, whether a claim or set of interrelated claims are valid in the context of an assurance case is, in part, conditional on whether they are accepted as reasonable justifications by those who are tasked with evaluating the assurance case⁹.

Let's look at another example. Consider the following section of an assurance case for the aforementioned chatbot, which the developers have produced for the procuring organisation:

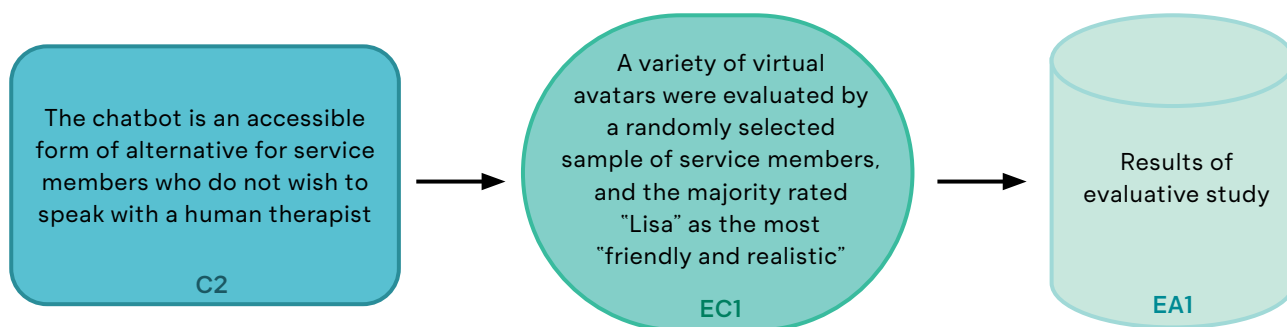


Figure 2.5: A portion of an assurance case for a chatbot¹⁰

As we have discussed, the organisation responsible for evaluating the trustworthiness of the AI system has to determine whether the evidential claim (EC1) is a reasonable choice to justify its parent claim (C2). They may, for instance, argue that EC1 is a reasonable claim, but nevertheless argue that it is insufficient on its own to justify the claim that the chatbot is an “accessible” alternative to human-led therapy. Alternatively, they may claim that it is not reasonable on the grounds that the ratings given by the service members are not relevant to establishing whether the chatbot is an “accessible form of therapy” but merely that the avatar is “friendly and realistic”. This example highlights a potential challenge associated with the development of assurance cases: determining what constitutes *relevant*, *sufficient*, and *reasonable* claims.

In the context of safety assurance, a large body of guidance has been established to help developers assess what claims they will need to establish and justify, and a key part of this guidance is the development of *argument patterns*.

What are argument patterns?

Argument patterns are starting templates for building assurance cases. They identify the types of claims (or, the sets of reasons) that need to be established to justify the associated top-level normative goal. **Figure 2.6** shows an example argument pattern for the technical goal of interpretability.¹¹

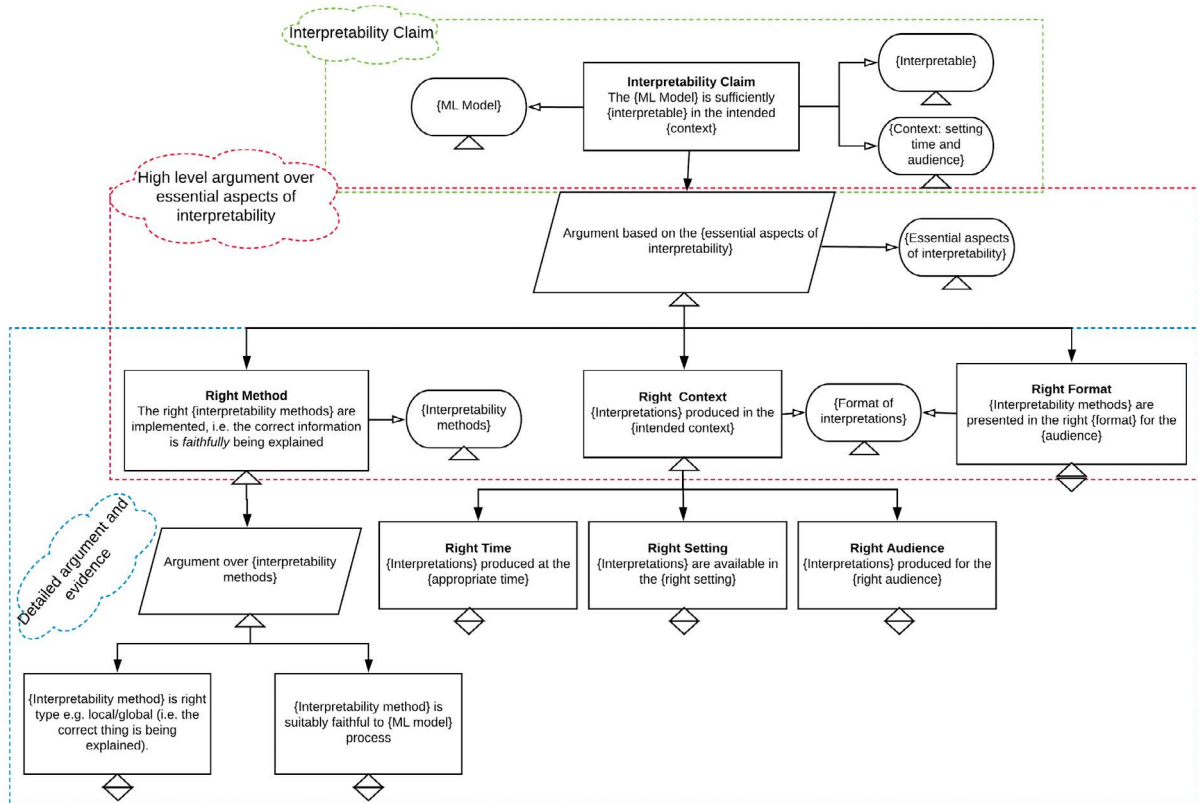


Figure 2.6: A pattern for an interpretability case (reprinted from Ward and Habli, 2020)

The pattern depicted in **Figure 2.6** shows a template for an assurance case that serves to justify the following top-level goal:


“The {ML Model} is sufficiently {interpretable}
in the intended {context}”

Here, the curly brackets serve as placeholders for specific variables that are properly established when a full assurance case is developed. A notable contribution of this pattern is the identification of three essential aspects of interpretability:

1 Right Method: The right interpretability methods are implemented, i.e. the correct information is faithfully being explained.

2 Right Context:

- › Time: Interpretations produced at the appropriate times.
- › Setting: Interpretations are available in the right setting.
- › Audience: Interpretations produced for the right audience.

3 Right Format: The interpretability methods are presented in the right format for the audience.

These three essential aspects subsequently serve to delineate the more detailed argument and evidential claims at the lowest levels.

Argument patterns, such as the one above, are helpful for the following reasons:

- › They provide a consistent and systematic approach for the reflective and deliberative activities carried out across a project’s lifecycle.

They speed up the process of developing assurance cases.

They provide reusable structures that, if used widely throughout a domain, could establish best practices.


But where do they come from?

Generalisable Patterns

In the case of the argument pattern from **Figure 2.6**, this pattern was proposed by the authors as a means to address a gap in the safety assurance of ML systems. As experts in their domain, and as a peer-reviewed contribution, this is a valid source for an argument pattern.

However, an alternative (though not entirely disconnected) means for achieving generalisable structures and patterns is through participatory engagement from stakeholders and affected users, perhaps building sample assurance cases and then extracting common themes. This is the method that we have explored in the current project on Digital Mental Healthcare and subsequently propose as a procedure for Trustworthy Assurance.

Much like ML algorithms, humans have remarkably effective (but biased) *pattern recognition capabilities*, some of which underpin our assessment and internalisation of ethical, legal, and social norms. As an example, James W. Nickel says of human rights:



“We can think of the emergence of a human right as the coming together of the recognition of a problem; the belief that the problem, is very severe; and optimism about the possibility of addressing it through social and political action at national and international levels.”¹²

Similarly, we can think of ethical and social norms as the shared recognition and subsequent externalisation of beliefs and attitudes towards events as diverse as acceptable etiquette during a dinner party through to permissible forms of punishment for various transgressions.

This understanding of the emergence of norms is crucial to ensuring the relevance, sufficiency, and reasonableness of evidence, and the legitimacy of corresponding trustworthy argument patterns.

In terms of the emergence of argument patterns, the three elements that we have explored already are, again, important: the top-level normative goal, the property claims, and the evidential support. Let’s take each of these in turn.

The phrase ‘trustworthy assurance’ creates a wide scope for top-level goals that may be deemed relevant to establishing trust (e.g. sustainable digital platforms, accountable methods of data governance, fair classifiers, and explainable decision support systems). As trustworthy assurance cases are developed for data-driven technologies, it is likely that we will see certain goals emphasised (and re-emphasised) over others. In turn, these normative goals will orient other projects and help cultivate best practices. In related work, we have proposed a series of ethical principles that have been developed to provide actionable insights and safeguards on responsible research and innovation in data science and AI.¹³ They are known as the SAFE-D principles:

- ➔ **Sustainability**
- ➔ **Accountability**
- ➔ **Fairness**
- ➔ **Explainability**
- ➔ **Data** (Quality, Integrity, Protection and Privacy)

These principles have been refined and validated in a wide range of domains, and were originally based on a broad understanding of the typical harms and benefits associated with data-driven technology (e.g. starting from the felt injustices or needs of users and stakeholders, and developing principles to reflect these challenges). Therefore, unlike alternative frameworks they are tailored to the specific needs and challenges of responsible, trustworthy, and ethical data science and AI, rather than, say, importing or revising existing frameworks such as biomedical ethics.¹⁴

However, the SAFE-D principles were designed to be domain-neutral starting points. That is, we did not presume that these principles would capture the ethical, social, and legal values that are dominant in digital mental healthcare. Instead, the present project undertook a process of exploratory engagement and participatory design to explore which ethical values and principles were relevant to the specific context of trustworthy digital mental healthcare, and whether specific SAFE-D principles captured these. We will return to this point in **Chapter 4** where we analyse our findings from the project's workshops.

Turning now to the property claims and supporting evidence, as assurance cases are communicated for specific goals we will likely see sets of property claims and supporting evidence used more than others as justifiable and accepted reasons for establishing the respective goal. For instance, as developers focus on goals like 'accountability', core attributes of the system and project are likely to be emphasised as relevant targets (e.g. constructing traceable data pipelines, establishing mechanisms to support auditing processes, ensuring accessible documentation).

Returning to the SAFE-D principles once more, we have previously developed a set of core *attributes* for each of the principles, which a) identify the types of properties that need to be established in a project or a system to ensure the relevant goal is obtained and b) the stages of the project lifecycle where actions can be taken to implement the respective property. **Table 1.2** shows an example of the core attributes for 'sustainability'.

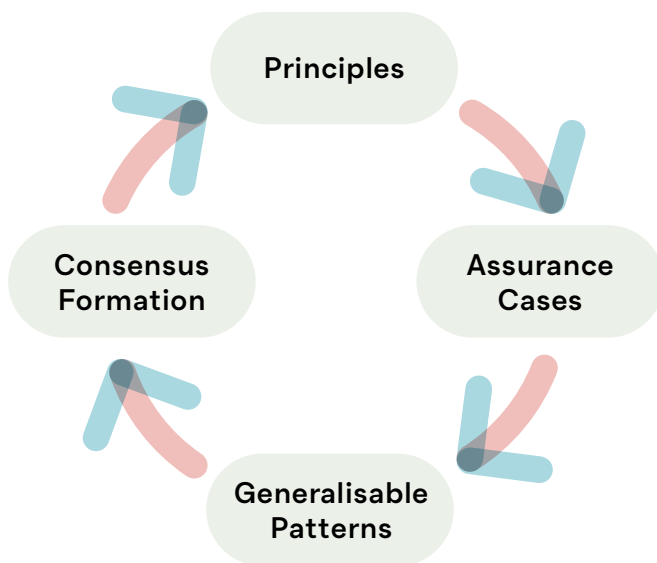
Table 1.2: A summary of the core attributes for the principle 'sustainability'

CORE ATTRIBUTE	DESCRIPTION
Safety	<p>Safety is core to sustainability but goes beyond the mere operational safety of the system. It also includes an understanding of the long-term use context and impact of the system, and the resources needed to ensure the system continues to operate safely over time within its environment (i.e. is sustainable). For instance, safety may depend upon sufficient change monitoring processes that establish whether there has been any substantive drift in the underlying data distributions or social operating environment. Or, it could also involve engaging and involving users and stakeholders in the design and assessment of AI systems that could impact their human rights and fundamental freedoms.</p>
Security	<p>Security encompasses the protection of several operational dimensions of an AI system when confronted with possible adversarial attack. A secure system is capable of maintaining the integrity of its constitutive information. This includes protecting its architecture from the unauthorised modification or damage of any of its component parts. A secure system also remains continuously functional and accessible to its authorised users and keeps confidential and private information secure even under hostile or adversarial conditions.</p>
Robustness	<p>The objective of robustness can be thought of as the goal that an AI system functions reliably and accurately under harsh or uncertain conditions. These conditions may include adversarial intervention, implementer error, or skewed goal-execution by an automated learner (in reinforcement learning applications). The measure of robustness is, therefore, the strength of a system's functional integrity and the soundness of its operation in response to difficult conditions, adversarial attacks, perturbations, data poisoning, or undesirable reinforcement learning behaviour.</p>

CORE ATTRIBUTE	DESCRIPTION
Reliability	The objective of reliability is that an AI system behaves exactly as its designers intended and anticipated. A reliable system adheres to the specifications it was programmed to carry out. Reliability is therefore a measure of consistency and can establish confidence in the safety of a system based upon the dependability with which it conforms to its intended functionality.
Accuracy and Performance	The accuracy of a model is the proportion of examples for which it generates a correct output. This performance measure is also sometimes characterised conversely as an error rate or the fraction of cases for which the model produces an incorrect output. Specifying a reasonable performance level for the system may also require refining or exchanging the measure of accuracy. For instance, if certain errors are more significant or costly than others, a metric for total cost can be integrated into the model so that the cost of one class of errors can be weighed against that of another.

Again, we are not proposing that these principles and core attributes should be adopted in digital mental healthcare as the respective goals, claims, and evidence. However, they could provide a starting point for the refinement of domain-specific principles while argument patterns emerge and become crystallised.¹⁵

Figure 2.7 offers a simple graphic to help visualise this process as it relates to trustworthy assurance.



These preliminary remarks about trustworthy assurance serve as a foundation for understanding and contextualising our project’s research and the recommendations we derive from our findings.

Figure 2.7: Process of consensus formation for ethical principles as constraints on trustworthy assurance



03

Applying Trustworthy Assurance—Digital Mental Healthcare at UK Universities

The University Context

Workshop Information

Analysis

- › Contextual Challenges
- › Methodological Challenges

Chapter Overview

This section is the first of two sections that present findings from research conducted with stakeholders and/or affected users. Specifically, this section presents findings from research conducted on the application of trustworthy assurance to the procurement of DMHTs for use in the higher education (HE) sector.

First, we provide an overview of the digital mental healthcare landscape at UK universities, detailing the pressures faced by university teams and the range of services currently on offer across the country.

Second, we present findings from a series of participatory engagements conducted with students and administrators at universities across the UK. We outline a series of contextual challenges to the ethical deployment of digital mental healthcare in higher education (HE) before exploring how the methodology of trustworthy assurance might be introduced in this sector to help tackle these challenges. For each challenge identified, future recommendations for the sector are provided.

Broader lessons for the ethics of digital mental healthcare are also fed forward into Chapter 4 where further recommendations for policy-makers and developers are given.



THE UNIVERSITY CONTEXT

Prior to the onset of the Covid-19 pandemic, the suggestion that student mental health was in crisis across UK universities was already prominent.¹ Media attention intensified around 2017, for instance, in response to a cluster of high-profile suicides.² Since then, concern has only grown further, and focus has turned to whether the crisis has worsened due to the increased social isolation brought about by Covid-19 and remote learning.³

While media reports have been criticised for their simplistic focus on suicide figures as a metric for student wellbeing, research by The Office for Students has found that lengthy waiting times for counselling and a rise in help-seeking behaviour have both put an increased strain on services. All this has occurred during a period where the overall student population has grown.⁴

Research continues to point to significant challenges facing university mental health services given factors such as the fivefold increase in students disclosing mental health conditions between 2007 and 2017⁵ and the lack of capacity to address student concerns quickly. For instance, NUS research found that only one in six students received professional support within one week of reaching out.⁶

In this high-pressure context, systemic changes have been proposed. Key policy documents, from the University Mental Health Charter to Universities UK Step Change Framework and IPPR's report on student mental health, have all called for a "whole university approach", for better cohesion between university departments, and for a greater focus on positive wellbeing. In addition to structural shifts, these recommendations have increasingly referenced digital interventions as a key tool within the student mental health offering.⁷

Digital interventions are frequently seen as a "natural step" for student mental health services as they expand beyond the traditional counselling model,⁸ while youth mental health is seen as a prime target for AI with the NHS AI strategy noting their first task with regard to mental wellbeing "is to look into children and young people's mental health" using AI-driven solutions.⁹

Amid these policy recommendations for modernisation, digital mental healthcare offerings designed for the general population have proliferated with The Office for Students suggesting 43,000 wellness and medical apps are now available for smartphone use.¹⁰

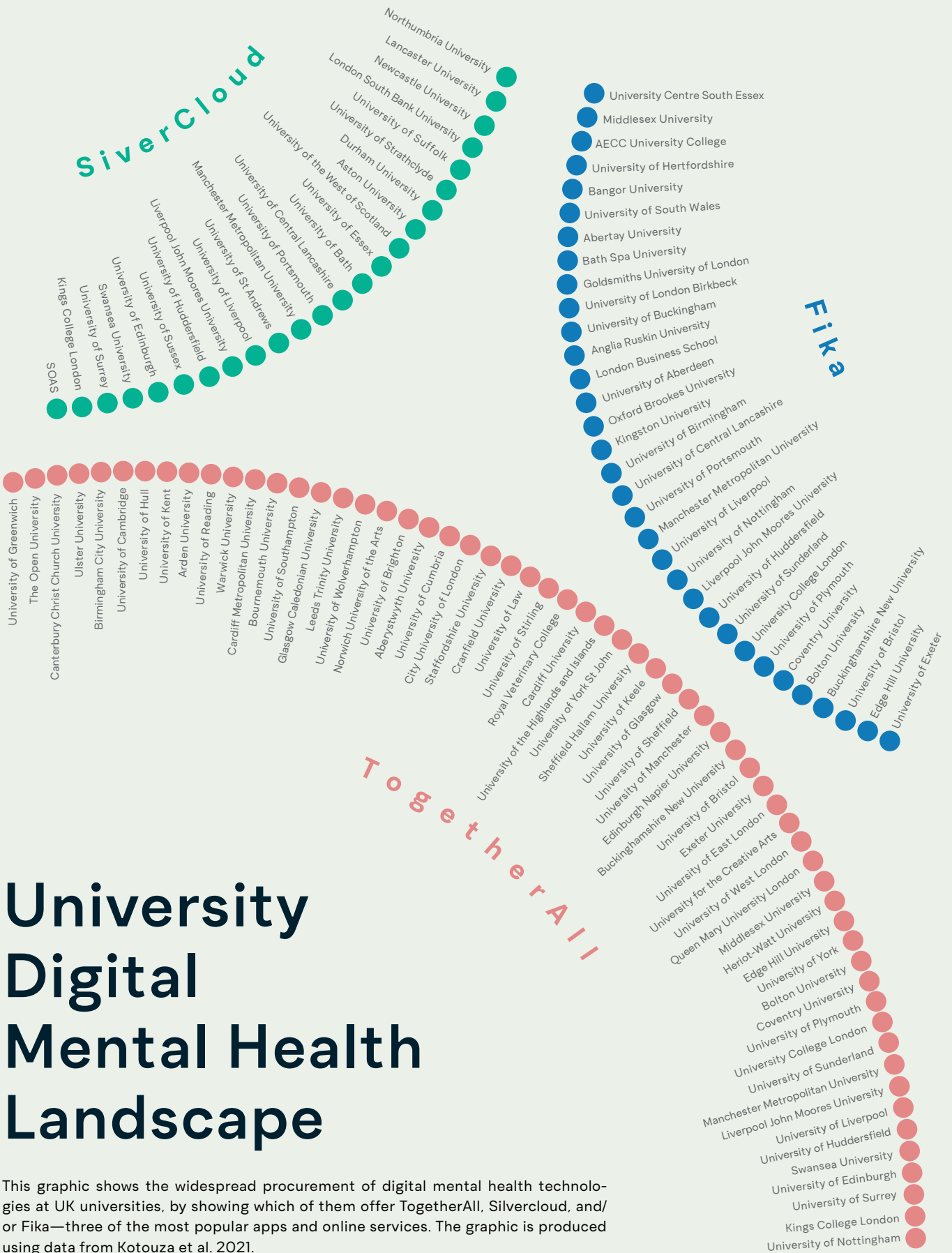
It is no surprise, therefore, that these new tools have appealed to universities facing rising service demand alongside tight budgets, and that digital mental healthcare adoption has advanced at pace across UK universities.

One way in which the growth in digital mental healthcare has been amplified is through universities' informal recommendations of external apps. For example, among the 24 Russell Group Universities in the UK, 15 recommend students try Headspace, eight point in the direction of Calm, and nine recommend SAM (Self-help App for the Mind), to name just a few of the apps listed on university webpages, typically under the heading of 'self-help resources'.¹¹

Another way in which digital mental healthcare has infiltrated the university sector is through the formal procurement by universities of digital mental healthcare tools designed with the specific needs of student populations in mind. "Kotouza et al. have mapped these shifts and found that, as of 2021, 'more than half of UK universities' have a contract with at least one from SilverCloud and Togetherall while 'Fika won over 35 UK university contracts between 2019 and 2020'".¹² (see full-page infographic below).

This rapid adoption of DMHTs may be seen as a perfect solution for the university sector, especially when the claims made by these digital service providers are taken at face value. For example, a co-founder of Fika has claimed that "mental fitness approaches like Fika's present a new, far more sustainable solution to the nation's mental health needs",¹³ while Kooth describe their platform as a "safe and confidential space to share experiences and gain support".¹⁴

However, it is important that universities are given the tools required to assess these claims themselves, and to do so using systematic, reliable, and well-validated procedures and standards. The [University Mental Health Charter](#), for instance, emphasises that the "need for quality assurance extends to other interventions, such as the provision of digitally based services". Further guidance is required for universities to be able to carry out such quality assurance, and in particular for university teams to effectively assess these digital tools on the basis of their ethical implications. Here, we present findings to inform universities in doing such due diligence.



University Digital Mental Health Landscape

This graphic shows the widespread procurement of digital mental health technologies at UK universities, by showing which of them offer TogetherAll, Silvercloud, and/or Fika—three of the most popular apps and online services. The graphic is produced using data from Kotouza et al. 2021.

WORKSHOP INFORMATION

To assess current challenges to the ethical and trustworthy deployment of digital healthcare mental technologies in a university context, we ran a series of participatory engagements to seek the feedback of university students and mental health teams.

The findings presented below are based upon a series of participatory engagement events conducted by The Alan Turing Institute between January and March 2022. These consisted of semi-structured interviews with university administrators leading mental health teams at 10 universities across the UK (henceforth 'administrators') alongside research workshops attended by 25 students enrolled in undergraduate and postgraduate courses at UK universities (henceforth 'students').

Table 3.1: Summary information about the workshops and interviews

GROUPS	> Administrator
PURPOSE OF INTERVIEW	<ul style="list-style-type: none">> To identify which data-driven technologies or services are used by higher education institutions to support or deliver on their duty of care for students.> To understand what metrics or properties are used to evaluate a data-driven technology or service prior to its procurement and deployment.> To determine whether the administrators and students share the same goals and values when evaluating a service.> To test the trustworthy assurance method and how it may be applied to specific case studies.
TYPE OF ENGAGEMENT	> Semi-Structured Interviews
MAIN ACTIVITIES	<ul style="list-style-type: none">> Exploring ethical principles and sharing current procurement approaches.> Understanding duty of care.> Assessing the trustworthy assurance methodology.

Table 3.1: Summary information about the workshops and interviews

GROUPS	> Students
PURPOSE OF WORKSHOP	> To explore a set of illustrative case studies that were designed to support the development of trustworthy assurance cases and identify significant ethical principles and values. > To build trustworthy assurance cases using a prototype platform developed for this purpose.
TYPE OF ENGAGEMENT	> Group Workshops
MAIN ACTIVITIES	> Which values and principles matter most to you? > Evaluating case studies.

ANALYSIS

Contextual Challenges

Key findings from these engagements conducted with university administrators and students are presented as follows. First, we explore contextual challenges to the ethical deployment of DMHTs at universities, as raised by students and administrators. Second, we present findings on how the implementation of trustworthy assurance to this sector may resolve challenges. In each case, distinct perspectives from administrators as compared to students will be highlighted and conclusions for the future of ethical digital mental healthcare procurement summarised.

As discussed in the [Introduction](#), digital technologies raise a distinct set of ethical issues when deployed in a mental health context. For example, it is necessary to consider their impact on the therapeutic relationship and the privacy implications of using sensitive health data.¹⁵ Many of these ethical issues continue to be relevant at UK universities. However, the challenges posed by emerging DMHTs are specific to the context in which they are deployed. Therefore, this sub-project set out to determine HE sector-specific challenges to the trustworthy and ethical procurement and assurance of DMHTs.

Contextual challenges, ranging from obstacles to transparency and accessibility to external pressures and institutional constraints, are summarised to inform key conclusions for the HE sector.

Challenge 1

Duty of care: a legal or ethical goal?

To determine which digital technologies to procure and provide, universities must first have a clear understanding of the overarching goal of their service. Defining this top-level goal also constitutes the first stage in developing a trustworthy assurance case.

In a university context, defining this overarching goal introduces its own set of challenges. Universities are understood to have a duty of care to their students' mental health. This has been defined by AMOSSHE—the Association of Managers Of Student Services in Higher Education—as the duty of the institution to “to deliver its educational and pastoral services to the standard of the ordinarily competent institution, and, in carrying out its services and functions, to act reasonably to protect the health, safety and welfare of its students.”¹⁶ Therefore, as part of this sub-project we explored whether duty of care could serve as a top-level normative goal for a trustworthy assurance case—going beyond the SAFE-D principles.

All digital technologies must support the university in delivering upon this duty. But, to assess whether a particular procurement choice is in-keeping with a university's

responsibility, more detail is needed to translate duty of care into a fully specified goal.

For university administrators, our engagement found that legalistic understandings of duty of care were common as references to the avoidance of negligence and to ensuring institutional competence and compliance were frequent. This is unsurprising, and one noted that, “if the worst outcome happens and you’re in a coroner’s court”, you must be able to show that due diligence has been done and procedure has been followed. Moreover, students also made frequent references to the importance of legal mechanisms in holding institutions to account.

Administrator Perspective

“We take the duty of care very seriously. And I think it is not the easiest thing to articulate. We talk regularly with our legal team about duty of care. If we are writing documents which reference the duty of care, we always talk with them about that.”

“In terms of duty of care, that brings in things like the equalities act and various educational laws and policies so that universities aren’t discriminating against students based on access or based on protected characteristics, and they are following all of the laws.”

Student Perspective

However, despite these legalistic framings, it is clear that the law does not provide a complete guide to action. AMOSSHE’s account on duty of care clarifies these challenges, noting first that “student law is still evolving”, thus creating “difficulty in providing a legal definition of an institution’s duty of care”, and second, that “there is a balance between what the university should do as a legal minimum and what they could do based on a university’s perceived moral obligation”. These uncertainties can cause confusion for administrators.¹⁷

Administrator Perspective

“I think again it is such a problematic area for higher education institutions. I’m not sure there is any clarity within the sector about that duty of care.”

Ambiguities with regard to the nature and scope of a university’s duty of care can be observed when universities must balance tensions between the preservation of student autonomy and the management of risk through forced intervention. Administrators raise the difficulty of these decisions on whether to intervene, illustrating uncertainty around duty of care.

Administrator Perspective

“They are 18, they are adults. If they don’t want our interventions, there is very little we can do. So, for us, it is balancing that fine line all the time.”

“I think we are at a very interesting period of somebody’s life. The complications are, obviously, over-18 somebody will be in 99.9% of cases an adult. So, understanding the duty of care you have to... people as children or minors is very different from this new stage. But there does seem to be some muddying of the waters with this and it comes up a lot.”

The everyday practices of a university mental health team must, therefore, go beyond the pursuit of compliance to grapple with ethical decisions on when specific interventions may or may not be appropriate. In short, legal understandings of duty of care must be supplemented by ethical values which guide the actions of teams in their selection of services.

For administrators, the ethical values raised as key to duty of care include: ‘inclusivity’, ‘putting the student first’, ‘consent’, ‘avoiding negative impact on the institution’, ‘best interest of every student’, ‘evidence-based decisions’, ‘benefit to all student groups’ and ‘value for money’. For students, key ethical values raised include: ‘transparency’, ‘inclusivity’, ‘anonymity’, ‘autonomy’, ‘privacy’ and ‘trust’.

While these ethical goals do show significant overlap, it is also clear to administrators that perspective is crucial to defining duty of care. First, different stakeholders define the goal of student mental health services differently. Second, perspectives on duty of care have shifted over time. Finally, the very introduction of digital services can influence who is seen to take responsibility for mental ill-health and so have an impact on how duty of care is understood.

Administrator Perspective

“Parents think we’ve got more responsibility than I think. Probably students sometimes think we’ve got less responsibility.”

“I think we are heading towards in loco parentis. Towards having that responsibility that schools have. I think that’s where the regulation is heading, if you look at things like the Student Mental Health Charter. All of the things are pushing us more and more that way.”

“Digitisation of mental health services may further the idea that mental health is an individual’s issue and responsibility, rather than addressing the collective mental health and structural causal issues around wellbeing.”

Student Perspective

Continued work is needed to understand if and how duty of care can be translated into a fully specified goal for a university mental health team and how this goal can inform specific actions during the procurement of DMHTs. A participatory approach, considering the views of all impacted stakeholders, will be essential to clarifying the nature and scope of this responsibility.

CONCLUSIONS AND RECOMMENDATIONS

Universities should reflect on their duty of care in both legal and ethical terms.

Shifting understandings and perspectival differences mean that a participatory approach to defining a top-level goal is crucial.

Challenge 2

Organisational constraints: defining roles and responsibilities

Ambiguities surrounding the relative roles and responsibilities of different teams and institutions to protect student mental health have also at times hampered effective decision-making with regard to digital mental healthcare. Descriptions of the decision-making rationale provided by administrators suggest a procurement process which is frequently reactive rather than proactive, with a tendency to stick to the status quo or shift tack dramatically during a crisis.

Administrator Perspective

“There was a general panic early March of ‘oh my god, what is going to happen?’”

“Quick leaps. It wasn’t particularly an evolution. We’d been considering it [Togetherall] ourselves for a couple of years so it was on the horizon but that [the pandemic] was the thing that really pushed us to do that.”

“With TogetherAll, a lot of universities have been offering it a long time, and there was a clinical trial, but anecdotally when you go on the mail base and talk to other heads of services and ask if their students are using it, they find students just aren’t really using it. That’s what we find. Paying a lot of money and people aren’t using it. But they don’t want to look bad or roll it back.”

Reactive decisions are in part a consequence of the lack of clarity surrounding who is responsible for decisions on the procurement of DMHTs. First, there are significant ambiguities

surrounding the relationship between university mental health services and the NHS. During engagements, administrators revealed that responsibility for procurement of digital services such as Togetherall resided in most instances with the university, but in certain cases with the NHS. In particular, administrators suggested that at times of crisis, where one institution was struggling to cope, the other had stepped in to fill a gap through purchasing a digital service. This dual responsibility can result in inconsistencies in who takes procurement decisions and in the long run could create a vacuum of responsibility if it is unclear where leadership lies.

Administrator Perspective

“That is the million-dollar question of where our services stop and the NHS starts and the piece around filling that gap because there is definitely a gap between.”

In addition, within the university, shifts towards a ‘whole university approach’ have occurred and were described by administrators. However, while greater involvement of departments such as financial services, accommodation, and examinations is taking place, less thought has been given to where responsibility for digital solutions may lie within this matrix and what capacity building is needed to develop the technical skills required to review digital technologies, monitor student use, and horizon scan for new directions. While procurement processes at many universities remain thorough in many respects, this results at times in decisions taken by clinical or legal teams, without dedicated attention to technical and ethical questions. In describing their procurement teams’ top priorities when reviewing digital technologies, administrators raised both clinical evidence and value for money. Other concerns specific to data-driven technologies and their ethical implications did not feature as prominently, indicating potential gaps in data literacy.

Administrator Perspective

“When I pull them [mental health team] in to look at new apps and things, what you find is they are trained practitioners, they are counsellors, but their digital skills are not up there.”

“I remember BigWhiteWall came around and did a slick presentation and then it was around the table who thinks this is a good idea. Rights, yeah, let’s go for it.”

Finally, as roles and responsibilities are clarified, a clear space for student involvement must be carved out. This was a priority raised by both students and administrators during engagements, but practical challenges must be addressed to involve all students and resolve possible tensions between student perspectives and clinical evidence. Additional challenges to student involvement in the procurement of digital technologies, such as the high entry costs sometimes associated with the evaluation of AI-driven tools, must also be addressed through education and collaboration between students and experts.

Administrator Perspective

“Universities like to claim they are giving the students what they want but a lot of the time they are not really consulting students. We often have to convince HR and other systems to talk to students.”

“Universities are often torn between do I go with the student voice or hard research. I will tell you which one they go with, student voice. The student voice doesn't have to be many, they just have to be loud. We are not talking about thousands of students having an impact, we are talking about vocal students or the students union.”

“Creating dedicated spaces for students to voice their ideas and needs combined with evidence that these were somewhat taken on board. Utilising relationships with students and personal tutors/student support. Strong links with student societies and other student intermediaries.”

Student Perspective

“Students need to be involved from the initial stages, so as to undercut serious biases working their way into the system. Furthermore, the biases within university cohorts should be examined – university intake of working class and state school students, POC, etc. is often not representative of the wider population, and technologies must make sure they are accessible to them as well, so that they are lasting into the future.”

CONCLUSIONS AND RECOMMENDATIONS

Clarity is needed on the relative responsibilities of the NHS and university teams during the procurement of DMHTs.

While university teams evolve to focus on interdepartmental cooperation, greater attention must be paid to assigning responsibility for digital offerings and ensuring technical skills have been developed.

Engagements with administrators suggest that compliance with data protection laws is well established while there remains a gap in organisational readiness with regard to other methods for reviewing algorithmic techniques. Cross-domain communication may be essential for these review processes, potentially including IT and research departments. Consequently, designating responsibility for digital technologies should be considered as part of the shift to a 'whole university approach'.

Clarity on roles and responsibilities can also help designate responsibility for horizon scanning and proactive research rather than reactive decision-making.

Meaningful student involvement is necessary and capacity building may be needed to facilitate meaningful contributions on technical topics.

Challenge 3

External Pressures

Greater clarity on responsibility for digital technologies within university mental health teams will only go part of the way to resolving current challenges. Significant pressures placed on mental health teams are outside of their control. In particular, concerns were raised throughout interviews and workshops surrounding the potential privatisation of student mental health.

Administrator Perspective

“While we are higher education institutions, we are businesses, and there is a conflict of interest with a business decision versus a clinically-driven, sound-evidence decision about what works and what doesn't.”

“For-profits don't tend to do things out of the kindness of their own hearts. So how much are universities paying for this? Will the students have to pay? Will it be full of ads? Where is the profit coming from?”

Student Perspective

The relationships between universities and developers do suggest for-profit companies are having a significant influence on the procurement of DMHTs. The constant flood of digital offerings creates an environment where mental health teams struggle to ensure services are driven by student demand. Additionally, the nature of pitches received from developers often fail to provide the evidence or information necessary to evaluate services. Finally, the volume of products on the market has led to a dichotomy whereby administrators either spend significant time filtering through proposals or ignore them altogether.

Administrator Perspective

“What I tend to get is lots of emails offering me a magic wand.”

“These apps are coming at you from every direction, and I must get at least three emails a week.”

“They [providers] are not telling me what I need to know.”

“I am absolutely inundated with emails every day from providers of these things and I just ignore them.”

These pressures which prevent services from being driven by student need are not only coming from private sector developers but from within the university sector itself. First, pressures from within an institution can lead to the adoption of a digital service on the basis of reputational concerns. Second, network pressures from comparisons with other universities can drive groupthink with regard to DMHTs.



Administrator Perspective

“I think it is about defensive practice as well. We just reviewed before the pandemic, we renewed BigWhiteWall, and I was quite keen to ditch and shift and look at alternatives. My boss at the time said I think we need to stay with it because it is a defensive action. If we get an FOI, it’s good to say all students get offered this. It’s worth 15k was the line I got, don’t care whether it works or not.”

“If you want to question decisions, you’re told ‘all these other Russell Group universities are offering BigWhiteWall or free subscriptions to X, Y, X’ so it seems that decisions are made more out of panic or what other institutions are doing because we need to look just as good.”

CONCLUSIONS AND RECOMMENDATIONS

Universities must work together to leverage collective influence on developers such that developers provide them with more detailed information on the ethical implications of new technologies. The Trustworthy Assurance methodology provides one way of doing this (See [Methodological Challenges](#)).

Networks such as AMOSSHE, the UK’s Student Services Organisation, can play an important role in stepping forward to provide evidence-based guidance on these topics and help to prevent decisions being taken on the basis of reputational comparison.

Challenge 4

Algorithmic aversion: diverted resources and prioritising human services

In a university context, where resources are constrained and digital solutions have been described as offering value for money, it is unsurprising that fears have been raised over technologies replacing in-person services, despite this going against the desires of both administrators and students. Determining when digital solutions are not appropriate, desirable, or morally permissible, therefore, constitutes a key challenge for the HE sector. Administrators note that despite their intention to supplement human services with digital, there have been times where more could be done to avoid the diversion of resources to digital technologies.

"We all know that universities are oversubscribed and when you are offered something like this, you feel as if you are at the bottom of the list, and your mental health and what you say is all part of criteria as to whether you are 'bad enough' or 'ill enough' for support. I think this is really damaging to someone's mental health and can worsen their state of mind."

Student Perspective

Administrator Perspective

"What I never want is that we see digital services as an alternative to coming in and seeing our team."

"We could probably be doing better to supplement and complement rather than replace."

Appeals to the importance of prioritising in-person services are significant. However, further specificity is required to determine how this can be achieved in a system where resources are constrained, and administrators inevitably face tough choices on where to direct funds. A more detailed understanding of students' and administrators' motivations for algorithmic aversion is necessary, therefore, to determine when and how digital technologies should be delivered. For administrators, concerns over clinical efficacy and the management of risks on platforms, as well as a lack of demand from students, formed the primary motivations for algorithmic aversion. In contrast, students focused on concerns around the dehumanisation of mental health, the lack of empathy offered by technology and the potential exacerbation of social isolation by digital offerings. Further research into these and further motivations for students' and administrators' algorithmic aversion is essential if student mental health services are to effectively deploy digital technologies.

Administrator Perspective

"What we want to know is has it been evaluated, what's the research that sits behind this and they [my team] have about zero tolerance of anything that just looks nice or doesn't have that really robust background."

“Students don’t want more apps and more digital interventions. They want a small group of evidence-based digital interventions that focus on positive outcomes, and they want to see staff face-to-face.”

“The dehumanisation of mental health. If students and staff come to see mental health as something solved by apps, the very complex nature of mental health can be undermined. We cannot automate mental health and wellbeing.”

Student Perspective

“As someone who has gone to therapy and other support groups, it is very beneficial to do the work and even asking for help and seeking out support and turning up to someone in person is very nerve-wracking and that part of the process is so beneficial to mental health as it helps you face some anxieties.”

“There is a certain level of isolation and problematic self-sufficiency that could be encouraged by directing users towards certain technologies.”

Taking these motivations into account can help to reveal a more complex picture where students and administrators oppose digital technologies for specific reasons and in specific circumstances. This detail is necessary to guide administrators in choosing when and how to deploy digital technologies, and when to stick with in-person services.

Administrator Perspective

“I think there’s slight value around students that have got really low-level concerns. So, sleep, procrastination, and all those things. I think some services offer great self-diagnosis. And then of course around how to do some of those things which actually, yeah, there is probably a benefits of being able to do that on your phone in the warmth of your own flat or whatever, without having to come in and see one of my advisors who might tell you exactly the same thing. So those are the two areas I see it. I don’t ever want us to get to a point where we see it as a solution to replace the ability to come in and see somebody, particularly for more risky students or students that are struggling.”

CONCLUSIONS AND RECOMMENDATIONS

In an environment where resources are constrained, careful consideration is needed over whether the decision to procure a technology is justified. Where resources are limited, justification of expenditure must go beyond claims that digital services do not replace in-person support to make transparent where funds come from and why the benefits of a digital service are seen to outweigh other options. This may also present a valuable opportunity to relative efficacy of traditional in-person mental health services.

Administrators' and students' reasons for preferring in-person services appear to differ. Consultation of multiple student and staff groups is, therefore, essential to determining when and why digital solutions may be inappropriate.

Challenge 5

Accessibility and fairness

In light of these varied motivations for aversion to digital solutions, it is clear that a more thorough understanding of different students' needs is essential. During engagements, key concerns were raised surrounding accessibility and fairness in digital mental healthcare, as many participants proposed a more tailored approach was needed. While increased accessibility of mental health services was described by both students and administrators as a primary advantage of digital technologies, a one size fits all approach has meant that access to effective care has not been improved for all groups.

Administrator Perspective

"There are certain student groups who would not like to go to one-to-one therapy where something online or an app or self-help, something more empowering that they can do in their own time. Something like that suits certain student groups. Broadly, with the little bits of research we have done, before the pandemic, it suits groups with social anxiety."

"It is assumed all students want the same thing, so we are just going to give them big white wall. Even if it is not being supported in any other way. It is just assumed students will have the motivation to use this app. But not all students are appropriate for self-help. You find even outside of technology, not everyone is suited to CBT."

"It [technology] can be more accessible, particularly for those with anxiety or mobility issues."

Student Perspective

"Providing accessible tools for 'basic' needs (e.g. productivity tools, which are not too concerning and can be widely used)."

These limitations of digital technologies' ability to improve accessibility need not be a reason against their deployment. In some instances, the ability to reach different groups, such as those who feel the stigma of reaching out in person, can be seen as a positive. However, to ensure accessibility is prioritised across the student population, a greater awareness of who is (and who is not) served by digital technologies is required, alongside an in-depth understanding of which student needs are currently unmet by in-person services. During engagements, some administrators suggested males were engaging more with digital technologies, some pointed to females, others to differences in engagement across academic disciplines and finally some noticed no patterns of engagement at all. Further research into this is needed for universities to move away from a one-size-fits all approach and design solutions which explicitly address the needs of all student groups.

Administrator Perspective

"Actually, if there was more work involving different student groups and not assuming every student is a first-year middle class white academic undergraduate and there are other groups that can benefit as well, that's where we can really learn."

For students, concerns about fairness and health equity were significant. First, concerns were raised around digital poverty and the importance of challenging assumptions frequently made about student populations (e.g., that all students have easy access to technology). Second, students expressed worries about biases within these technologies and how algorithmic bias can impact their experiences. Here, students advocated for specific consideration of minority student groups with the possible designation of 'safe spaces' on such platforms based on protected characteristics or allowing students to actively choose who they were speaking to in order to avoid potential bullying or harassment.

"Digital poverty is a thing. Just because you are at university, that doesn't mean you have a smartphone, a laptop, or even a private space where you can access the platform."

Student Perspective

I hate to be devil's advocate, but I don't think everybody has access to devices. I do get the point around accessibility, but I think companies or universities also need to be offering the support which will lead to them accessing these services and closing the digital poverty gap."

"Information can be taken out of context and the cultural nuances with people from a range of backgrounds can be missed."

During discussions on the accessibility and fairness of digital technologies, proposals were made by students for accessibility to be understood in more subtle ways whereby digital solutions are not simply rolled out across the student population, but instead tailored to students

with specific needs. It was also proposed that humans should mediate access to technology where possible so that a student can be matched with a service that helps them at the level they are ready for. Finally, empowerment of student choice was prioritised by participants in order to ensure students have control over the care they receive.

CONCLUSIONS AND RECOMMENDATIONS

Greater integration with research departments is needed to ensure evidence is available on who is best served by specific mental health services. Universities can draw on their own research resources in order to take a deeper look at who is and is not benefiting from digital interventions.

Resource allocation within student services should take account of issues such as digital poverty, especially in light of the rise in remote learning following the pandemic.

Algorithmic bias is a pervasive issue beyond the mental health sector and one which administrators must be aware of so that thorough questions can be asked of service providers in advance of service roll out.¹⁸

A balance must be struck between empowering students to make their own choices without placing the burden of responsibility for mental health on them.

Challenge 6

Transparency and communication

Both students and administrators raised the importance of transparent and accessible communication about what services are available to students. Both groups also recognised that there is a lack of cut-through in current communication strategies leading to students being unaware of which services are on offer at their institutions.

Administrator Perspective

“[Communications] can be quite difficult and especially around this sort of issue, around support. Through the pandemic, the availability of support has probably increased accessibility because a lot of it’s been online. And yet you still come across students who say, ‘there is no support, I don’t know where to find it’. And you kind of think, well there was a student newsletter, there’s all these social media posts across multiple channels, the student union that talked their language are talking about this stuff and TikTok is messaging about this, but still you haven’t picked it up? So, it is quite a challenging environment.”

“I do not have adequate awareness of what is available.”

Student
Perspective

“I think they should promote them more so I would actually know what is available.”

“I know that I get it a lot at the bottom of an email or a newsletter that I don't tend to read and I know that lots of people don't read.”

However, among students, the need for more awareness on what services are available was not the first priority regarding transparency and communication. Rather than focus simply on advertising services to students through multiple channels, students emphasised the importance of universities communicating the limitations and personal costs associated with these technologies to them, such as the loss of privacy or limited clinical efficacy. Such transparency is also emphasised as important to facilitating informed consent.

“There is an asymmetry of knowledge as users don't know/consent to their data being used for these purposes. The business model is not communicated to the user and is exploitative.”

Student
Perspective

“All these policies will say things about sharing information or data with 'trusted individuals', 'trusted organisations', 'trusted researchers', but there is nothing about what makes these organisations trusted. And you can't expect somebody in crisis to actually pour through all the terms and conditions.”

CONCLUSIONS AND RECOMMENDATIONS

To ensure students are aware of the services available, communications should be delivered by as many stakeholders as possible to include academic staff, student unions and student societies. These communications should also incorporate considerations around accessibility (e.g. alternative formats).

University mental health teams should be careful to communicate honestly with students about both benefits and risks of technologies, including transparent communications about efficacy and data sharing.

While in-depth and legally binding privacy policies are essential, an accessible breakdown of key information (e.g. through FAQs or video messages) is crucial to 'informed' consent.

Methodological Challenges

Overall, feedback from administrators and students suggests there is a long way to go to ensure responsible digital mental healthcare innovation and procurement across the UK HE sector. Progress can be made by following key recommendations set out here for universities. However, significant challenges will remain so long as changes do not take place elsewhere in the digital mental healthcare ecosystem. Systemic changes from developers, policy-makers, and university leadership teams are needed to leverage the collective power required to foster a responsible innovation landscape.

The trustworthy assurance methodology provides one route through which such a culture of responsible innovation and transparent communication could be facilitated so that university stakeholders are given all necessary information to review new mental health tools offered to them. Due to the potential of this methodology in this sector we asked administrators to provide feedback on its deployment. Crucially, feedback indicated that this methodology would fit within current procurement practices, provide structure for the critical assessment of developers' claims, and encourage more transparent communication among university staff about the rationale behind procurement decisions.

Administrator Perspective

“My initial reaction is it’s bloody brilliant. I’ve not seen anything like that and if something like that landed in my lap I would seize upon it.”

“What I’m looking for is a really quick way of working out whether this is something that’s worth looking at or not.”

“People like me are not naturally going to be flowchart kind of system development people. So, something like that [the assurance methodology], that can help us ask those questions and take these people through their paces, because what we get is very slick presentations, would be very helpful.”

“I think it would be really helpful to have a standard set of expectations that, you know... a standard expectation of having to demonstrate what they were trying to achieve.”

Nevertheless, several challenges to the effective deployment of trustworthy assurance, as raised by university administrators and students, must first be addressed. These are set out below in order to inform conclusions in Section 3 for the future of the assurance methodology.

Despite this positive feedback, the methodological challenges set out below must be addressed for trustworthy assurance to be deployed in the HE sector.

Challenge 7

Interpreting assurance cases

Trustworthy assurance cases (as set out in [Chapter 2](#)) can appear complex. This introduces two key challenges for administrators tasked with assessing whether a particular assurance case meets their requirements for the procurement of a new DMHT.

First, technical skills are required to assess such cases and ensure evidence matches with property claims. Second, assurance cases cannot be treated in isolation and must often be evaluated alongside clinical standards. It is important to avoid treating clinical efficacy and ethical standards as equivalent. The assurance exercise is not intended to provide a universal stamp of approval that ethical concerns have been addressed. Instead, it will require contextual interpretation on whether the evidence provided by a developer is sufficient for deployment in a specific context.

As a result, despite placing the greatest burden of responsibility on developers, this methodology does require administrators, as key decision-makers in the HE sector, to have capacity to effectively review assurance cases. Policy-makers must, therefore, take on responsibility for ensuring the necessary knowledge and organisational readiness to assess digital technologies is provided to universities. This could be offered through consultation with national bodies such as AMOSSHE, The Office for Students or Universities UK or through dedicating resources to capacity building within each higher education institution.

Challenge 8

Uptake by developers

To obtain necessary evidence and compile it into an assurance case requires time and resources on the side of developers. A key obstacle to the deployment of trustworthy assurance therefore rests upon the motivation of developers to do this due diligence. Currently, such motivation will be minimal as standards across the sector place few demands on developers.

Nevertheless, the collective purchasing power of universities across the UK is significant and can be leveraged in order to place further requirements on developers. This may be done through standards setting at the national level, for example through networks such as AMOSSHE. Additional pressures can also be placed on university networks by organisations such as the National Union of Students whose collective voice can influence procurement decisions.

In setting high standards for developers, there must also be efforts not to exclude smaller mental health providers from having the resources to produce assurance cases. For this rea-

son, the burden placed on developers may be reduced by sharing best practices or building repositories of publicly accessible assurance cases and argument patterns which can be used as a starting point for developers as they embark on the ethical reflection process. These routes towards cross-sector best practices are discussed in greater depth in [Chapter 5](#).

Challenge 9

Use as a communication tool

Finally, in a university context, concerns were raised regarding the utility of trustworthy assurance in communicating to students that due diligence has been done and ethical implications have been considered.

In conversation with the students, we discussed whether there was perceived value in administrators sharing assurance cases directly with the student body. Although there was no negative response, concerns were addressed about the complexity of assurance cases. Similar concerns were raised by administrators. In brief, stakeholders thought assurance cases represented too much information for effective communication, and that a summary would be preferable. Access to the full case could be made available for any students who were interested in order to facilitate both accessible communication and full transparency.

The header features a light pink background with various geometric shapes: a large dark blue 'D4', several squares and circles in shades of pink and light green, and a set of concentric circles in the top right corner.

D4

Co-Designing Trustworthy Assurance—Stakeholder Engagement

Workshop Information

- › Participants
- › Methodology and Activities

Analysis

- › Workshops (1a and 1b) with policy-makers and regulators, developers, and researchers
- › Workshops with users of digital mental health technologies

Chapter Overview

This section introduces and analyses the findings of several stakeholder engagement events, which were conducted to a) identify participant's attitudes towards DMHTs, b) understand which ethical values and principles they view as significant, and c) explore how to use this information to construct trustworthy assurance cases and argument patterns for relevant ethical goals.

First, we introduce the objectives, structure, and content of the workshops.

Second, we analyse the findings of our workshops, drawing connections with the methodology of trustworthy assurance. These findings support the development of two argument patterns, presented in chapter 5, which serve to distill recurring themes from all of workshop participants that were deemed significant.

Finally, we offer several recommendations for policy-makers, regulators, and developers, based on the preliminary results of the project.



WORKSHOP INFORMATION

Participants

In the previous chapter we discussed our engagement with university students and administrators. This work was treated as a sub-project because of the specific focus on the HE sector as a limiting context. In this chapter we address workshops with a wider range of stakeholders and from a broader perspective. While the objectives remain the same across these two chapters, the findings and analysis in this section are representative of a wider range of concerns.¹

The stakeholder groups we consider in this chapter are as follows:

-
- 1** Policy-makers and regulators in healthcare

 - 2** Developers of DMHTs

 - 3** Researchers working in disciplines adjacent to digital mental healthcare

 - 4** Users with lived experience of DMHTs

Representatives from the first three stakeholder groups were invited to participate in a series of two workshops, the first of which laid the foundation for a subsequent participatory design workshop.

In contrast, users of DMHTs (4) were invited to a separate workshop (offered either online or in-person), which was organised with and facilitated by the McPin Foundation—a mental health research charity that provide advice and support on research strategies to involve participation and expertise from individuals with lived experience of mental health issues. This decision was made to ensure that participants were fully supported by experts throughout the workshops, and that our analysis of the findings was further supported by domain experts.

Methodology and Activities

Full details of our methodology and activities are provided in [Appendix 1](#) (available on our website). Summary information is included in **Table 4.1**.

Table 4.1: Summary information about the two sets of workshops	
WORKSHOP	> 1a
GROUPS	> Policy-makers and regulators, Developers, Researchers
PURPOSE OF WORKSHOP	<ul style="list-style-type: none"> > To introduce participants to the methodology of trustworthy assurance > To identify key ethical values and principles that were salient or significant in the evaluation of digital mental healthcare
MAIN ACTIVITIES	<ul style="list-style-type: none"> > Introductory presentations on a) the current landscape of digital mental healthcare, including representative harms and benefits, and b) the methodology and purpose of trustworthy assurance > Group discussion exploring the ethical values and principles associated with the design, development, and deployment of DMHTs, using case studies developed by our team
WORKSHOP	> 1b
GROUPS	> Policy-makers and regulators, Developers, Researchers
PURPOSE OF WORKSHOP	<ul style="list-style-type: none"> > To explore a set of illustrative case studies that were designed to support the development of trustworthy assurance cases > To build trustworthy assurance cases using a prototype platform developed for this purpose
MAIN ACTIVITIES	<ul style="list-style-type: none"> > A group discussion of the chosen case study (voted for by participants in the previous workshop) to ensure familiarity with the relevant details of the case study > A participatory design activity in which the participants collectively develop an assurance case for a specific ethical value or principle (e.g. health equity, explainable decisions)

WORKSHOP > 2

GROUPS > Users of DMHTs (in-person; online)

PURPOSE OF WORKSHOP > To identify participants attitudes towards digital mental healthcare in general, and salient ethical issues more specifically.

MAIN ACTIVITIES

- > Exploratory discussion on the possible harms and benefits of digital mental healthcare.
- > Identification of key ethical values and principles.
- > Evaluation of sample claims made by a hypothetical team about actions or decisions undertaken during the design, development, and deployment of DMHTs.

ANALYSIS

KEY FINDINGS

1

—

All groups emphasised fairness as a key ethical principle, but the specifics of how fairness was understood differed between groups.

2

—

Additional emphasis was placed on ethical priorities that could be captured by either the accountability, explainability, or data SAFE-D principles (e.g. informed consent, transparency).

3

—

Goals that are not directly coupled to any specific ethical principle², such as clinical efficacy, were nevertheless significant topics for consideration among regulators and developers.

4

—

Ensuring sufficient understanding of the trustworthy assurance methodology proved to be challenging in the time available. This was the case even with the participants who attended two workshops, where the first included preliminary material on the methodology.

As with the workshops described in the previous section, we conducted thematic analysis on the findings from the workshops, with the goal of addressing the objectives set out in the Introduction.³

In the following sections, we first discuss the specific themes for each set of workshops and then explore cross-cutting themes and differences, expanding on the Key Findings section above.

Workshops (1a and 1b) with policy-makers and regulators, developers, and researchers

SUMMARY

- › Nearly all of the ethical issues raised could be easily captured by the SAFE-D principles and their core attributes. However, additional space and emphasis is needed to capture the following concepts: choice, patient choice, self-determination, autonomy.


- › Fairness was prioritised by the majority of participants. The principles was strongly linked to considerations such as access to services, unequal distribution of health outcomes across demographic groups, bias in algorithmic decision-making, and diverse and inclusive participation in service design.

- › Participants expressed positive sentiment towards the trustworthy assurance, noting its perceived value for processes such as transparent auditing, assessment, or procurement.

- › Producing assurance cases was a challenging exercise for many, but there were no signs that these barriers could not be addressed with additional user guidance and familiarity.

Workshop 1a

The first workshop (1a) ensured that participants had sufficient information about the trustworthy assurance methodology, which was required for the second workshop. This information was provided while minimising the likelihood of priming the participants to evaluate our case studies with reference to specific ethical values or principles, such as the SAFE-D principles. Therefore, there were fewer findings from workshop 1a than with workshop 1b.

 **Note:** Our goal was to identify which ethical principles mattered most to them, so we were careful not to highlight that we had already developed an existing framework (SAFE-D principles) that could unduly influence their feedback.

However, one relevant activity from workshop 1a that is worth mentioning was the explicit request to identify and discuss ethical values and principles that were seen as salient or significant in the context of the design, development, and deployment of trustworthy DMHTs.

The following word cloud shows participant answers for the question,



'What values and principles matter to you?'



Figure 4.1: A word cloud displaying answers to the question, 'What values and principles matter to you?'

Immediately, it can be seen that **transparency**, **privacy**, and **evidence-based** were clearly significant for our participants. And related concepts, such as **accountability**, **explainability**, and **clarity** are also salient.

However, there is also a wide variety of terms here, and many are either synonymous or closely related. For instance, **self-determination**, **autonomy**, **informed consent**, **control of own data**, and **patient choice** could be clustered together. And so could **equity**, **fairness**, **equality**, and **equity of access**.

Although intended as an exploratory and preliminary activity, the findings represent a useful source of contextual information that can help illuminate some of the themes that emerged during the workshop discussion and activities. For instance, the vast majority of the concepts map onto our existing framework, and overlap with the SAFE-D principles and corresponding attributes (see **Table 3.2**).⁴

Table 4.2: Mapping the word cloud concepts onto the SAFE-D principles

PRINCIPLE	CONCEPT
Sustainability	evidence-based, fit for purpose, safeguarding, monitoring, cost effective, follow up, redress, safety, usefulness, impact
Accountability	transparency, safeguarding, accountability, expert led, regulated, monitoring, honesty, redress, monitored closely
Fairness	accessible/accessibility, fairness, co-designed, compassion, bias, equity, co-produced, diversity, equality, equality of access
Explainability	transparency, evidence-based, clarity, accessible, honesty explainability, monitored closely
Data Quality, Integrity, Privacy and Protection	privacy, control of own data, regulated, honesty, usefulness, safe and secure, confidentiality

Two considerations can be extracted from this mapping:

- 1 The SAFE-D principles provide an informative starting point for ethical reflection and deliberation in digital mental healthcare, as they do in other domains, and would likely serve as useful normative goals for trustworthy assurance cases.
- 2 There are gaps and nuances in the framework when applied to digital mental healthcare that need to be addressed.

In terms of the second consideration, there are a few clarifications that need to be made.

Firstly, the main gap relates to the ability for the principles to capture concepts such as **choice**, **patient choice**, **self-determination**, and **autonomy**. The appearance of these concepts is not surprising. Patient autonomy, informed consent, and participatory decision-making in healthcare are longstanding ethical values, and are reflected in well-known bioethical principles.⁵

In the original domain-general setting in which the SAFE-D principles were designed, informed consent and autonomous decision-making were captured under principles such as fairness

and explainability (e.g. ensuring that information about an algorithmic decision is accessible and explainable to users). However, as we will see shortly, there are nuances in the design, development, and deployment of DMHTs that put pressure on the choice of subsuming these values into another principle.

Secondly, there are other principles, such as **human-centred**, **human rights**, **honesty** and **closed loop systems** that are either ambiguous or do not fit cleanly into the existing framework. In the context of the first two, this is simply because they stand outside of the SAFE-D principles as meta-frameworks (e.g. human rights law). For instance, the SAFE-D principles have been put forward as a means for safeguarding human rights.⁶ However, in the case of **honesty** and **closed loop systems**, there was simply insufficient information during discussion to determine whether these are merely outliers or express an existing attribute that fits within the framework.

Fortunately, the activities and discussion from the second workshop help emphasise more salient topics that were deemed significant by the stakeholders.

Workshop 1b

The second workshop (1b) focused on a participatory design activity that was created to evaluate the trustworthy assurance methodology and attempt to operationalise some of the ethical principles explored in the first workshop.

For the main activity, participants were asked to review and discuss the ethical issues related to a specific case study and then develop a hypothetical assurance case that communicated how a set of decisions or actions had been undertaken to justify the ethical goals and properties that they had discussed. The groups were free to choose the goal. And the case study, which had also been selected by participants, involved the use of a decision support system that offered tailored and real-time recommendations to a psychiatrist during consultation with a patient (e.g. during assessment) (see [Appendix 1](#) on our website).

Two breakout groups were formed and the (incomplete) assurance cases are depicted below.

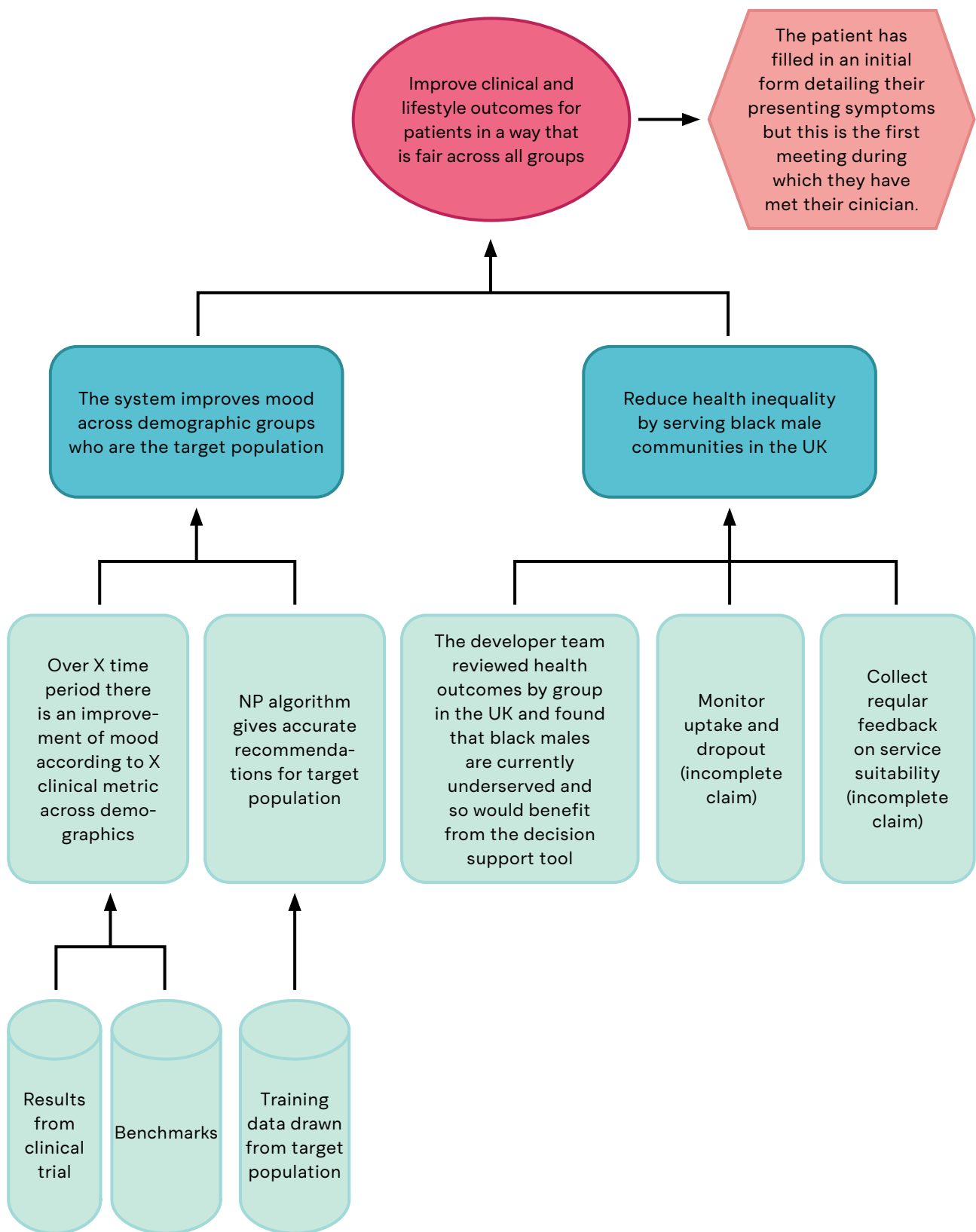


Figure 4.2: The assurance case for breakout group 1, focusing on ensuring fair outcomes for patients.

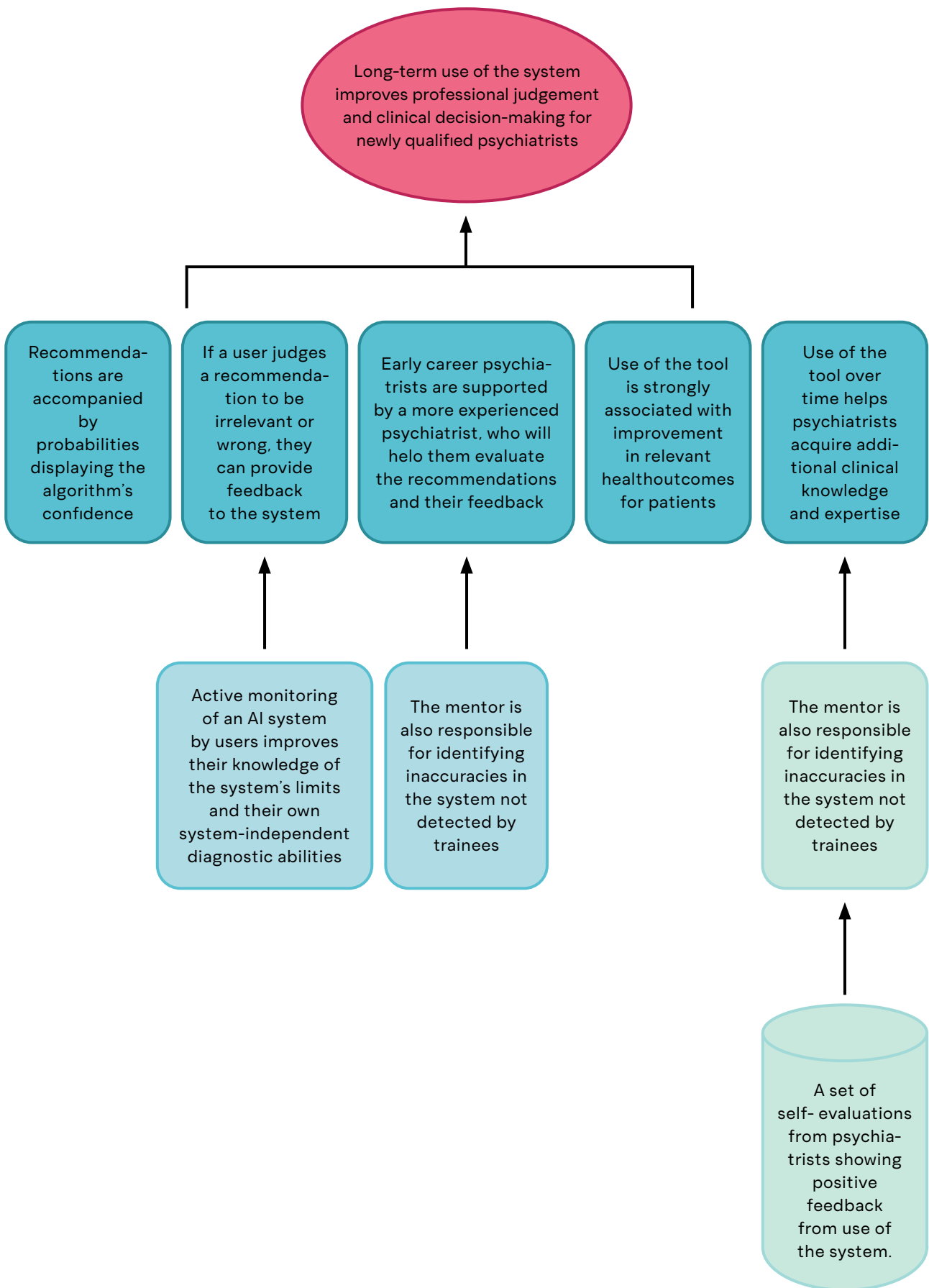



Figure 4.3: The assurance case for breakout group 2, focusing on supporting the professional judgement of psychiatrists.

As noted above, fairness was a significant focus for the participants, and so it is unsurprising that one of the groups chose to explore an assurance case related to this goal.

The assurance cases are incomplete because a lot of time was spent in discussion. However, some of the statements made during discussion can help elucidate aspects of the case. For instance, the choice to emphasise a group that is underserved, was linked to aforementioned components of fairness such as unequal access:



“There are clearly real benefits for many people being able to access digital technology but there are also many people who won’t be able to access it. Prior to even thinking about the introduction of digital technology in mental health services, we were aware of very longstanding inequalities in access to an outcomes from mental health services. There’s a fear, particularly during COVID and lockdown and the move to ‘digital by default’ that some of those divisions will grow larger as a result.”

Interestingly, the second breakout group chose to focus on the impact of the hypothetical decision support system upon the healthcare professionals.⁷

As the image above shows, their goal was framed in terms of supporting the “professional judgement” of the psychiatrist. On its own, this goal would be underspecified, making it difficult to accurately link to any of the SAFE-D principles or core attributes. Fortunately, the discussion and property claims of the assurance case help clarify the intentions of this group, but some residual ambiguity remains due to the inclusion of two core themes:


-
- 1** The long-term effects of the system on professional development and judgement
 - 2** The responsibility of early career psychiatrists, who may be less able to challenge or contest automated recommendations.
-

These two themes emerged from discussion about the potential harmful impact of the system on the judgement of psychiatrists, such as the possibility of automation bias (i.e. the tendency for users to be unduly influenced by automated decision-making, even when their own judgement would be preferable or more accurate). For instance, when the group considered which forms of evidence they would expect to see included to provide assurance

that this risk had been managed and mitigated, they added an evidential claim and artefact that communicated additional mentor support (from a senior healthcare professional) and positive self-evaluation from the user (see right of **Figure 4.3**).

The inclusion of this segment in the assurance case cannot be divorced from consideration about the responsibility that a user has for their own decisions, as well as the institutional mechanisms of accountability that ought to be put in place to support users of decision support systems. Therefore, it remains unclear whether the goal of this assurance case could be captured by a single SAFE-D principle. On the one hand, Sustainability would be a good candidate for capturing the long-term impacts of the system on user autonomy and professional judgement. On the other hand, Accountability would be preferable for those attributes concerned with responsible decision-making and institutional accountability. These questions were not raised with the participants, nor is there sufficient information to infer their intentions.

However, the following quotation from one of the (developer) participants is illuminating for the challenges involved in ensuring responsible decision-making in situations where multiple organisations are involved in a distributed project lifecycle:



“Ultimately, once you hand over the software and you set them [procuring organisation] up, how they actually use it and why they’ve got it can be a bit of a mystery. Sometimes, organisations are looking for ways to support people but they don’t have much resource and they think digital might be a good way of doing that. So, maybe it’s still a good motivation but there’s also this expectation that digital can do a lot more than it can, and it’s a cheap way of ticking a box.”

In general, it was challenging to develop a full assurance case for several reasons:

-
- 1** Time limits imposed by workshop
 - 2** Lack of familiarity with the Trustworthy Assurance methodology
 - 3** Challenges of reconciling broad range of perspectives to co-create a shared goal and common understanding

The first two barriers would be easy to reconcile. For instance, although we dedicated a significant portion of time to introducing and exploring the trustworthy assurance methodology, more hands-on experience with the tool could have helped the participatory activity of developing an assurance case for the hypothetical case studies.

The third barrier, however, is harder to overcome. In our initial project planning meetings we considered running separate workshops for all of the stakeholder groups, but settled on mixed engagement events because of the perceived benefit of facilitating communication between different groups—a key benefit of the methodology itself. A portion of this barrier could be resolved with additional time, but the value-based and translational gap that typically exists between different groups (e.g. regulators and developers) will remain. Therefore, the following recommendation is proposed as a measure to address this challenge:



Readiness, skills, and training should be prioritised both within organisations (e.g. how to implement ethical considerations into the project lifecycle) and across organisations (e.g. how to develop and adopt best practices). In addition, common capacity building should be supported by regulators and industry representatives (e.g. shared risk mapping, regulatory gap analysis, and horizon scanning activities to help create and maintain a common pool of expertise).⁸

Feedback

Following the participatory activity, participants were asked to complete an anonymous survey, which was designed to elicit additional information about the perceived value of the trustworthy assurance methodology. Therefore, despite the small sample, it is worth analysing the responses before we turn to the final workshop with users of DMHTs.



To what extent do you agree/disagree with the following statement: “The methodology of trustworthy assurance would be helpful in identifying potential ethical risks which arise while designing, developing and deploying a DMHT”.

OPTIONS	#RESPONSES
Strongly Agree	4
Agree	9
Undecided	2
Disagree	0
Strongly Disagree	0

This initial feedback is positive. The majority of respondents 'agree' or 'strongly agree' with the statement, indicating support for the methodology despite the challenges faced during the activity.

Unfortunately, the sample size is too small to infer anything meaningful about the distribution of participants across these categories. For instance, whether developers were more positive than regulators or vice versa.⁹



To what extent do you agree/disagree with the following statement: “The methodology of trustworthy assurance would be a helpful means through which to communicate to other stakeholders that a DMHT is trustworthy”.

Similarly, the majority of respondents 'agree' or 'strongly agree' with the above statement, reinforcing our prior assumption about the communicative value of trustworthy assurance. However, as we will see shortly, there are some future areas for improvement, which likely explain the increased number of 'undecided' responses, which were primarily from policy-makers.¹⁰

OPTIONS	#RESPONSES
Strongly Agree	4
Agree	8
Undecided	3
Disagree	0
Strongly Disagree	0



What do you consider to be the primary advantages of the ethical assurance methodology?

The feedback from this question can be summarised as follows:

- › The primary advantage is having a capacity to support a structured and end-to-end approach to project governance, facilitated by shared aims and objectives for broader normative goals (e.g. health equity).

For instance, as one participant noted,

- › “Its [i.e. trustworthy assurance] emphasis on structure and evidence supporting claims which derive from an overarching goal. It really helps to be forced to think in these terms to keep an open mind about what requirements a certain system has at different stages of design, development and deployment. I love how flexible the system is, so that it can account for many technologies and contexts on the market.” [Researcher]

Although trustworthy assurance is a structured process, as this participant emphasises, flexibility is also maintained by enabling myriad goals, properties, and evidence to be selected to fit the context of a specific project. Where this flexibility is used to facilitate bidirectional decision-making about project aims and objectives, trustworthy assurance can (as another participant notes) support the development of

- › “A level playing field expectation from procurers, and a common improvement in practice.” [Developer]

And, in turn,

- › “will help developers, commissioners and service providers to know that equality/ethics has been considered.” [Policy-maker]



What do you consider to be the primary disadvantages of the ethical assurance methodology? Please specify any aspects of the methodology which you believe require further refinement.

While it is encouraging to see many positive responses, it is also important to consider critical feedback, as it is here that potential gaps and barriers can be addressed. The first critical comment attenuates the positive feedback about the methodology’s flexibility noted above:

- › “The open-ended nature of the methodology makes it difficult to know what to include in scope and may pose a difficulty for comparing assurance cases between manufacturers, for instance.” [Developer]

This is a valid concern, but can be addressed through the use of a) argument patterns and b) guidance about how ethical principles can be operationalised throughout the project lifecycle (see below and [chapter 5](#)).

The second point relates to organisational culture and readiness, and the challenge of considering competing incentive or disincentive structures—a theme also raised in the previous chapter:


- › “There are lots of examples of people producing equality impact assessments¹¹ that are little more than a tickbox exercise. It’s important they are produced, but it’s even more important they are of a high quality.” [Policy-maker]

We have previously emphasised that trustworthy assurance should not be reduced to a mere checklist or compliance exercise. The model of the project lifecycle is one means for mitigating the risk of misuse in this manner, as it emphasises the iterative and dynamic process of building an assurance case over the course of the entire project lifecycle. As such, the act of building an assurance case is not rendered a checklist exercise that is carried out at the end of a project as an afterthought, but is rather approached as subject to ongoing review. However, this prescription is going to be in conflict with alternative interests, as one participant notes:

- › “Assurance may be not in the interest of profit” [Researcher]

At present, our methodology is not supported by a theory of change for how organisations can adopt the methodology into their own practices. One possibility would be to work with public sector organisations and regulators to establish a requirement for those responding to tenders to provide an assurance case for a relevant ethical goal (e.g. non-discrimination, explainability). Alternatively, we could investigate how specific goals in the context of healthcare could expand upon existing regulations (e.g. medical device approval).

The following three questions were posed to the respective participants as a means of eliciting more specific feedback about perceived obstacles to the successful adoption and integration of trustworthy assurance into their respective practices.

 **As a developer, are there any objections or external obstacles which would prevent you from producing an ethical assurance case during the design, development and deployment of a new DMHT?**

The first two comments relate to a similar concern about organisational readiness and incentive/disincentive structures raised above:

- › “External obstacles are business needs and drivers that might cut down on the time needed to use a methodology like this properly. It can be hard to create and argue for time to give ethics proper consideration when there are business deadlines.”
- › “To produce an ethical assurance case would require additional effort. It would ideally be integrated into other early design processes alongside clinical safety case, data protection impact assessment, etc. The process would need to be usable in an (agile) product development context, where an ethical assurance case would be updated as changes are made to the tool in an iterative fashion.”

Here, our analysis and response echoes some of the comments made earlier (e.g. conflict with profit incentives). However, one additional comment can be made: having tailored versions of the project lifecycle model, which reflect the unique needs and challenges of specific domains, would help developers identify which actions could be undertaken (and when) to meet the goal of an assurance case. This development would, furthermore, create space for the development of supporting standards, as acknowledged by the following participant:

- › “Clarity on expected standards would be the largest obstacle. We noticed the biggest upswing in GDPR policy uptake came when we started offering standard starter policies.”



As a researcher working on DMHTs, do you see any obstacles to the uptake of ethical assurance in the sector?

The responses from researchers were mostly positive in the sense that few obstacles were identified. However, there was a skepticism about the likelihood of the private sector adopting the practices of trustworthy assurance.

- › “Ethical regulation, in the private sector, is non-existent. Theories around mental health, in general, are too underdeveloped, the digital context and the methods to use for research unexplored, non-rigorous, and misaligned with gold standards. Ethical assurance has no tangible rewards for a designer with a purpose or aim, independently of the positive principles used to design the system.”

We can, again, note that targeting procurement practices in the public sector may be a positive first step, and, moreover, that complementarity with existing regulation could increase adoption. This latter point was highlighted by the responses from the policy-makers.



As a policy-maker, do you consider the methodology of ethical assurance to be compatible with existing regulatory mechanisms in the sector? Please describe any obstacles to the adoption of ethical assurance in the digital mental healthcare sector.

- › “I’d recommend that you align to existing regs where possible (e.g. goals could be ‘conform to GDPR’ or ‘meet regulatory requirements’ rather than more abstract items) and make it easy for non-experts to use.”

The response from another policy-maker, however, suggests that caution should still be exercised when seeking alignment, in order to avoid confusion:

- › “It [trustworthy assurance] is compatible but it may overlap considerably with other requirements eg. ESF, DTAC etc. creating confusion and overload for developers.”

The final section of the survey, asked participants to offer any remaining feedback. Here, the selection of responses serves as a useful summary on the above analysis.

First, our analysis shows that while many recognise the value of the methodology, significant obstacles remain in the successful adoption of trustworthy assurance. Most notably, the friction presented by antagonistic incentive/disincentive structures. While our comments suggest possible avenues that may counteract some of the disincentives, a comment raised by one participant suggests that more work needs to be done to better communicate the positive value of using trustworthy assurance:

- › “How do we ensure developers of mental health tools use or are even aware of these types of methodologies? What is the incentive to use these methodologies?” [Developer]

We outlined many of the potential benefits of trustworthy assurance in the first section, but ensuring others are convinced of these values will take time. In alignment with the final two comments from our participants, a means for reducing the friction would be to build out additional case studies or examples of best practice, which can help serve as a point of orientation for other developers and organisations.

- › “The worked example made things easier to understand and think about.” [Developer]
- › “Having a best practice example to learn from and emulate would help our practice.” [Developer]

This suggestion is a variation of the recommendation above about supporting readiness, skills and training, and common capacity building. However, in the [next chapter](#) we will build on this by setting out a clearer proposal and recommendation for the incorporation of argument patterns.

Workshops with users of DMHTs

SUMMARY

The workshops with users exposed wide-ranging, nuanced, and interconnected attitudes, while contributing to practical and complementary recommendations for developers and regulators.

Four central themes emerged from the workshops:

-
- › Distrust as a barrier to accessing and using DMHTs

 - › Stakeholder and user engagement as a means for ensuring accountability

 - › Explainable technology and systems as a pre-requisite for informed choice

 - › Ensuring fairness by reducing digital exclusion, bias, and discrimination, and promoting social justice

While all of the themes are interconnected, the fourth theme especially is inseparable from the others.

Overview

The workshops with users of DMHTs were co-organised and facilitated by the McPin foundation—a mental health research charity. This ensured an additional level of support from those with domain expertise, in addition to the participants' lived experience, and helped reduce interpreter bias in our analysis.

We held two workshops (one in person and one online) to improve accessibility for participants (e.g. reducing geographic restrictions, supporting those who were uncomfortable/unable to attend in-person to still participate). The information that participants received and the activities that were carried out were identical across the workshops, except for the medium in which the activities were conducted (e.g. use of a collaborative whiteboard in the online setting).

There were two activities that participants contributed to. Both were designed to maximise the ability of the feedback to shape and inform the design of our methodology and recommendations while minimising the need for prior reading (e.g. information about argument-based assurance). A talk preceded each of the activities to ensure that participants were equipped to contribute in a meaningful way.

Activity 1: participants were asked to reflect on a range of possible use cases for DMHTs and evaluate possible harms and benefits by answering the following questions:

- › Which ethical values or principles matter to you in the context of digital mental healthcare?
- › What are some positive use cases for DMHTs?
- › What are some negative use cases for DMHTs?

Activity 2: participants were given a set of claims made by a fictional development team about one of the four case studies, and asked to evaluate the claim based on the following criteria:

- › Whether the claim appeared to be motivated by or support an ethical value or principle.
- › Whether they found the claim reassuring or whether it raised concerns.
- › What evidence, if any, they would expect to see to support or validate the claim.

In both activities, participants were encouraged to explore tangential points in dialogue with the group. The purpose of these activities was to provide a general scaffold for discussion, from which salient and significant themes could be identified with the participants. Therefore, in our analysis we do not differentiate between the findings from the two activities, but instead group them together and make specific recommendations linked to the relevant themes.

However, one output from the first activity can be presented as a stand-alone output. **Table 4.3** presents a summary of the positive and negative uses of DMHTs, as judged by the workshop participants. Although this feedback is incorporated into our own thematic analysis, the reader may find it illuminating to consider the responses prior to reviewing our subsequent analysis.

Table 4.3: Participants' perceptions about the positive and negative uses of DMHTs.	
POSITIVE USE CASES	NEGATIVE USE CASES
Useful for opening dialogue with clinicians	Predatory "targeted marketing", business/finance models that take advantage of people, using people's data to attain information and target them. Manufactured empathy is not empathy.
Tools for self-management and self-help	Limitations of the technology leading to problems. Difficulty in determining when there is an emergency, not understanding tone of voice.
Useful for remaining anonymous	Selling people's personal data
Could be useful for preserving continuity of care, and personalisation of care	Privacy issues (young people especially)
Useful for dangerous or violent person where clinical contact is bad/not recommended	Stalking, coercive control and abuse, people pretending to be identities that they are not.

POSITIVE USE CASES	NEGATIVE USE CASES
Potential for a deeper understanding of mental health difficulties due to the amount of data that could be collected	Infiltration in creepy ways into personal/sex life.
Accessibility - can do in your own time and from your own home. Useful for those who live remote areas where travel is not possible or expensive	Constant monitoring by the device, increased paranoia, over-reliance on device.
Reducing the load on psychiatrists	Increased loneliness. Social isolation can be exacerbated when you are talking to a chatbot, or you can become reliant on the chatbot.
Ability to share recovery (or other mental health) narratives digitally	Misrepresentation of outcomes. When usage time is monitored, not using the device can indicate deep depression or apathy but could also indicate that there are other good things going on in the real world they are engaged with.
	Tech use and being online has been problematic for some service users and giving it all up has been a good factor in their recovery.
	Some people don't all have access to privacy to use digital technologies in their own home.
	Need to keep up-to-date hardware to access may impact those on lower incomes and increase the electronic waste problem.


Thematic Analysis

The following themes were identified across the two workshops and activities. They have been co-developed by the users, the facilitators from the McPin foundation, and ourselves.

Distrust as a barrier to access and use

There were high levels of distrust and skepticism within the group regarding the societal and individual benefits of digital mental health. However, the sources and targets of the distrust or skepticism were nuanced and wide-ranging.


Several participants, for example, were keen to acknowledge that the issues with digital technologies should be set against the backdrop of the current issues facing mental health services (e.g. long wait lists, insufficient funding). This includes a recognition of the difficulty of getting face-to-face appointments and the biases of human healthcare professionals:



“I didn’t like it [online CBT], I was just desperate to have any form of counselling and because the waiting list was two years, I thought it was better than nothing.”

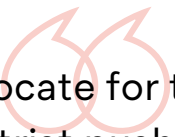
“You get people in the NHS who are as bad as chatbots. They may as well be robots.”

Some participants linked this distrust to cultural attitudes they held:



“My parents are [information redacted] and really value privacy and that is why I didn’t use a smartphone for a long time. And a lot of my friends who are African or Caribbean or Asian don’t have a smartphone because of privacy.”


Whereas others linked the source of distrust to potential conflicts of interest:



“Who is the advocate for the technology?
Is it the psychiatrist pushing it because it
makes their life easier?”

“Who holds the purse strings?”

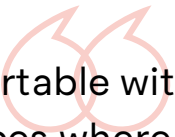
For some participants, the distrust or skepticism was directed towards specific technologies such as chatbots:



“[Chatbots are] good for customer service,
but sometimes it feels like they are being
used to replace humans.”

“How is it ethical to develop a solution like that, and
allow these technologies to exist on the marketplace
when they are doing more harm than good?”

But, again, the skepticism that participants held was typically nuanced and voiced with caveats:



“I am very comfortable with tech. There are
some circumstances where I trust technology
more than people. I trust an iPad food ordering
system than a human.”

In some instances, such as AI-assisted surgery,
involvement of technology should optimise
for safety. But in mental healthcare, human
interaction and involvement will always be vital
to a trustworthy and supportive relationship.

From these preliminary remarks, it is important to remember how we disentangled the concepts of 'trust' and 'trustworthiness' back in [Chapter 1](#). To recall, 'trust' can refer to a belief or attitude that is directed towards an object, person, or proposition (among other things), whereas 'trustworthiness' refers to the perceived property or attribute which an individual uses to determine whether to place trust (e.g. whether to trust a news article based on its quoted sources). Differentiating these terms is helpful for evaluating whether there are reasonable (and unreasonable) grounds for placing trust. For example, a person may have reasonable grounds for their distrust in an organisation, where the organisation has a history of violating data protection and privacy laws. In contrast, another person may have unreasonable grounds for their skepticism about the clinical efficacy of a technology based on an accessible, well-validated, and reliable evidence base.

Identifying the reasons for why users may trust or distrust a DMHT can help organisations assess and evaluate both the trustworthiness of their teams and services, and identify opportunities for intervention. Phrased as a recommendation:

RECOMMENDATION

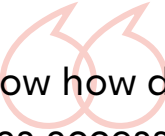
Organisations should consider both the trustworthiness of their products and services, but also the reasons why users may trust or distrust them. Four central themes emerged from the workshops:

Acting upon this recommendation requires organisations to engage stakeholders, which links to the next theme that emerged during discussion.

Accountability through engagement

In her BBC Reith Lectures, 'A Question of Trust', moral philosopher Onora O'Neill argues that, "we need more intelligent forms of accountability, and that we need to focus less on grandiose ideals of transparency and rather more on limiting deception."¹²

O'Neill's prescription captures many of the concerns and aspirations of the participants. For instance, several participants voiced concerns with the deceptive practices of some organisations to exploit vulnerable users through social media marketing (e.g. adolescents). Other participants viewed the over-reliance on privacy policies to be an instance of deceptive practice, as such policies rarely provide sufficient information to address a user's concern, such as data use:



“people should know how data is being used,
who has access to it.”

In contrast, participants were keen to express their desire for genuine forms of accountability and responsibility exercised through the life cycle of a DMHT, achieved through meaningful engagement and participation of stakeholders. The slogan, “nothing about us without us” comes to mind here¹³. And, there are also close ties between this theme and the subsequent one (i.e. explainability):



“No transparency without
accountability and explainability.”

This emphasis on engagement and meaningful forms of participation will be returned to, as it was a recurring and cross-cutting theme. However, several practical recommendations can be offered here in connection with the theme of accountability:

RECOMMENDATIONS

Accountability should be built into all stages of the project lifecycle, and requires both stakeholder engagement and also diversity within the project team (especially neurodiversity).

Where there is a risk of harm to users, organisations should be transparent about how these risks were identified (e.g. who was involved in the risk assessment), how they were mitigated, and what mechanisms for redress are available to impacted individuals.

Explainability as a pre-requisite for informed choice

As has already been noted, there was a strong dislike and distrust towards the perceived over-reliance on privacy policies. As several participants noted:



“The culture of small-print is pervasive, but the culture of consent is not built into a business model.”

“it is not consent because you are not informed.”

“Pooled permissions are a risk to privacy and there should be more modular options to accept or deny the Terms and Conditions. Usually, you must ‘accept all’ for an app to work.”

However, the following question from one participant inverts the perspective and shifts focus onto the user:

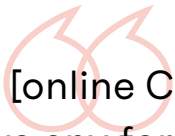


“How do we communicate our privacy policy.”

Traditional frameworks in biomedical ethics link informed consent to values such as patient autonomy. In short, without sufficient knowledge about how a service operates or the risks associated with a medical intervention, a patient has little to no meaningful choice about whether to engage or accept a recommendation from a healthcare professional.

The above quotation is a succinct and cogent way of capturing the ethical importance of these values. But its emphasis on a more active form of communication goes beyond the practical goal of informed and autonomous decision-making, and reiterates the importance of stakeholder participation as a form of meaningful input and influence.

To understand why this shift in framing matters, consider the fact that for many users there may be no practical choice about whether to engage if there is only a single option available to them. This may be because of long wait times, limited provision of services, or perhaps because a user is only choosing to engage at a point of crisis.




“I didn’t like it [online CBT], I was just desperate to have any form of counselling and because the waiting list was two years, I thought it was better than nothing. It will keep me from having suicidal ideation.”

Capturing these themes and building on the previous sections recommendations, we can add the following recommendation.

RECOMMENDATION

Information that is necessary to and supportive of informed choice should not be hidden within obscure privacy policies; it should be made accessible to users as explanations of how a system was designed, developed, and deployed. In doing so, organisations should be clear about how they define and operationalise key terms, such as 'mental health' or 'well-being' and how their understanding of the terms may have impacted the design, development, and evaluation of a service.

But the shift from choice to active involvement is not just about improving explanations to support participation, it is also about improving access more generally. As one participant acknowledges, this is fundamentally a matter of fairness.



“If everything is moving towards digital, who is going to be excluded. Is it going to be harder to access face-to-face care?”


This brings us to our final theme.

Fairness: reducing digital exclusion, bias and discrimination, and promoting social justice

Despite being left until the end of the chapter, this final theme stood out as one of the most significant and resonates with many aspects of the themes above and also with the other workshops.


In a similar manner to the other workshops, the plurality of concepts referenced in this theme's heading reflects the breadth, nuance, and interconnectedness of the ideas that were raised. For instance, the participants' understanding of what we call 'fairness' in the SAFE-D principle framework was nuanced and multifaceted. Although we are unable to capture all of the comments raised, the topics discussed touched upon centuries-old forms of sociocultural and structural discrimination or oppression, historical abuses of vulnerable groups by scientific research groups and institutions, epistemic injustices, and power imbalances that disproportionately affect marginalised communities.

Digging deeper into some of these topics, concerns about the negative impacts of digital exclusion and the widening digital divide, exacerbated by growing socioeconomic inequalities, were highlighted frequently. Some participants linked their concerns to gaps in current regulation and legislation:




“Ensuring inclusion and accessibility requires going beyond protected characteristics: disability & class and access to technology.”¹⁴

While others emphasised structural forms of exclusion in technology design:



“When designing based on AI and machine learning, we look at what works for the mass and the smaller minority communities and the rare types of people/personality are excluded by design.”

Similarly, some participants raised questions about the possibility of algorithmic discrimination due to varying levels of efficacy across demographic groups:



“How will it [the hypothetical NLP algorithm in one of our case studies] account for regional words and dialect? Slang terms? Cultural terms? Accessibility in different languages.”

These critical comments and considerations should not be isolated from the remarks outlined in previous themes or the following recommendations raised by participants:

- › *Distrust as a barrier to access and use*: historic forms of oppression, injustice, and discrimination partially explain why some individuals and groups have low levels of trust towards these technologies and the organisations that design, develop, and deploy them.
- › *Accountability through engagement*: the risks of harm and the likely benefits associated with DMHTs may not be shared equally by all groups. Inclusive stakeholder engagement is one mechanism by which oversight and accountability in the risk management process can be achieved. To paraphrase one participant, ‘those on a design team should be a diverse, invested group, and diversity should not be tokenistic’.
- › *Explainability as a pre-requisite for informed choice*: as a bioethical principle, ensuring informed consent is often linked to the ethical value of self-determination. The significance of the principle is understood by many to arise from a universal right to autonomous decision-making in matters relating to one’s health and well-being. While the domain of mental healthcare places restrictions on this right when it conflicts with other duties (e.g. protecting others from harm), these are limiting cases for which there are existing norms and guidelines in place¹⁵. In general, ensuring that an individual has sufficient access to the explanations needed about how a technology operates, in order to make an informed choice about whether to use the technology, has already been acknowledged as a vital ethical goal. However, there are many barriers in place to achieving this goal, and where they disproportionately affect certain groups (e.g. those with low levels of digital literacy or access to support services) this goal connects with the related goal of promoting social justice.

Improving health equity is already a key priority across many organisations in the UK.¹⁶ However, the longstanding impacts and challenges of COVID-19 are still affecting society, often in a disproportionate and unjust manner, and many lessons still need to be learned and adopted, as highlighted in [**Build Back Fairer: The Covid-19 Marmot Review**](#), from the Institute of Health Equity and Health Foundation.

Unlike the other themes, we do not offer any specific recommendations on this topic beyond the reiteration of the importance of stakeholder engagement and meaningful participation. This is partially because there is already a wealth of extant research produced by organisations across the public and third sectors offer evidence-based policy recommendations. However, it is also because we pick up on this theme directly in the next chapter and present an argument pattern to help promote the goal of fairness in digital mental health.



05

Developing Trustworthy Assurance—Argument Patterns for fairness and Explainability

Co-designing argument patterns

- › Fairness
- › Explainability
- › Evidential Considerations

Chapter Overview

In this chapter we present two argument patterns (i.e. starting templates for building assurance cases) that identify the types of claims, or the sets of reasons that need to be established to justify the associated top-level normative goal.

The first pattern is for assurance cases that aim to justify the fairness of a DMHT. The second is for cases that address the explainability of systems.

We also discuss relevant legislation, regulation, and best practice guidance that support and motivate the development of these patterns.



CO-DESIGNING ARGUMENT PATTERNS

While the assurance methodology is a tool in its own right, there is a missing component that was highlighted in [Chapter 2](#): argument patterns.

You may recall that argument patterns are reusable structures that serve as starting templates for building assurance cases. They identify the types of claims (or, the sets of reasons) that need to be established to justify the associated top-level normative goal. And, in doing so, they set useful constraints on both the deliberative process and evidence-generating and evidence-gathering exercises. We say more about the evidential generation and selection process towards the end of this section.¹

Before this, we present and explain two argument patterns for use in the assurance of DMHTs. The first is for assurance cases that address and justify the fairness of a DMHT; the second is for cases that address and justify the explainability of systems.



Why 'fairness' and 'explainability'?

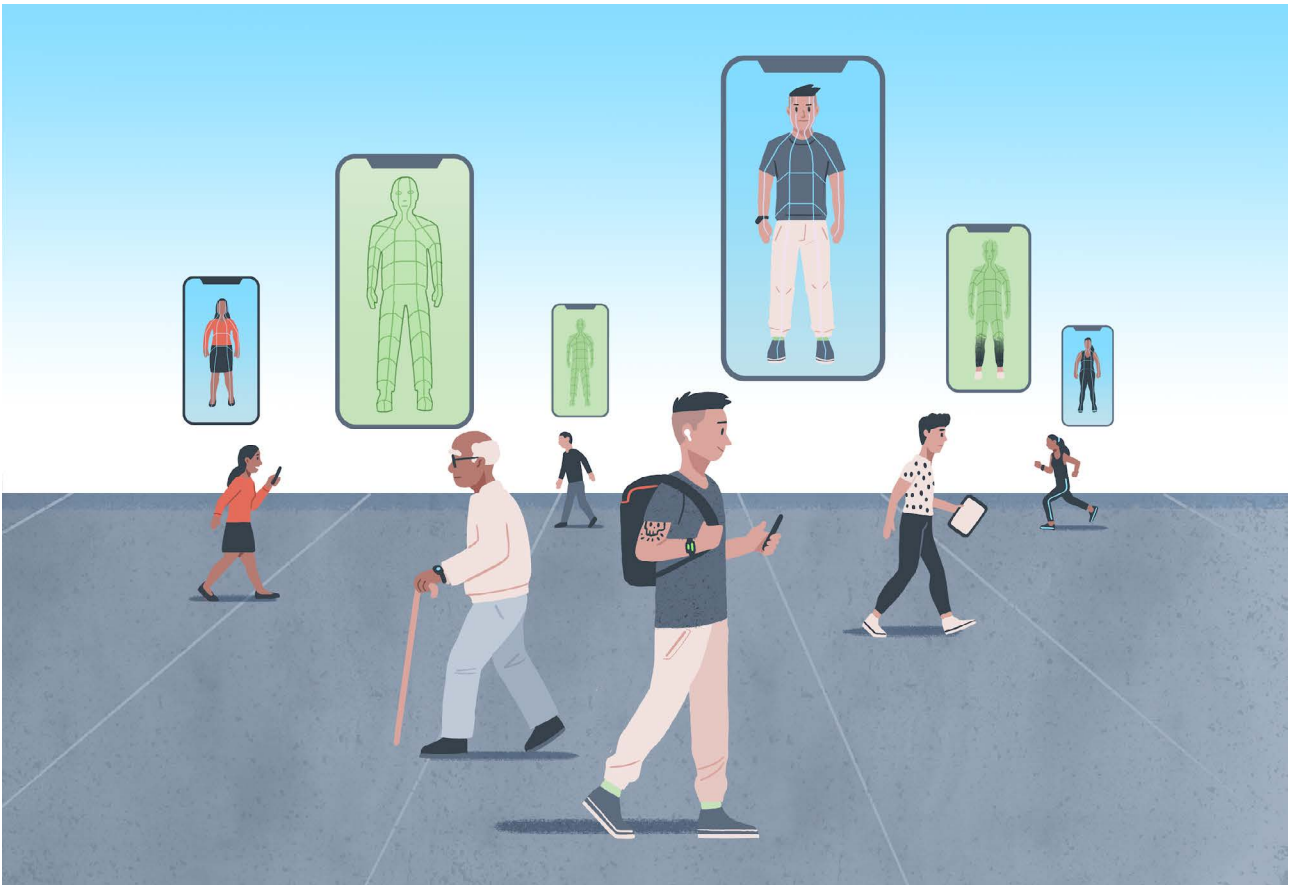
Our decision to focus on these two patterns is motivated primarily by the desire to capture the significant themes raised in the workshops. As such, the inclusion of a pattern for fairness is an obvious choice based on the discussions in the [previous chapter](#). However, the inclusion of one for the goal of explainability requires additional clarification.

Many participants in our workshops focused on ethical issues that could map onto multiple SAFE-D principles. For instance, considerations around transparency were sometimes raised in connection with mechanisms for holding organisations accountable and at other times raised in connection with a requirement to ensure users had access to information to support informed consent and decision-making—captured either by sustainability or explainability. Our second argument pattern is framed in terms of explainability as an attempt to be maximally inclusive of these wide-ranging considerations.

To recall, argument patterns in trustworthy assurance are always starting points for participatory forms of reflection and deliberation. They provide greater specificity than the top-level goal would on its own, and help operationalise ethical principles within the project lifecycle model. But they are no substitute for embedded processes of inclusive stakeholder engagement—in fact, they depend upon stakeholder engagement processes for their completion. Furthermore, they should not be used as a mere checklist for compliance.

Finally, our focus on 'fairness' and 'explainability' should not suggest that other patterns are not desirable or important. Rather, the co-design and development of additional patterns, including those that go beyond the SAFE-D principles are left for future research (see [Conclusion](#)).

Fairness



In the context of data-driven technologies, a core attribute of fairness is the equal distribution of risk and benefit across all groups of affected users. For instance, a technology that was highly accurate for users aged between 18-30 but became decreasingly accurate for older individuals would be unfair to those from higher age brackets.

The differentiation between risk and benefit leads to a subsequent distinction between corresponding duties or obligations (e.g. legal duties). Where there is a duty to ensure that a particular group of users are not exposed to disproportionate risks of harm, this can set up a so-called “negative duty” (e.g. a duty for a developer to not discriminate). However, negative duties often set only the minimal ethical standards. In other words, one can build a non-discriminatory service that does not harm anyone, but similarly benefits no one or benefits only a small portion of users.

Therefore, corresponding “positive duties” also exist to promote beneficial outcomes, sometimes focusing on those who are most disadvantaged—a so-called “prioritarian” approach to ethics (i.e. prioritising those who need the most support). A good example of this duality is the Public Sector Equality Duty, which sets the following objectives for all public authorities:

- › eliminate discrimination, harassment, victimisation and any other conduct that is prohibited by or under the Equality Act 2010;
- › advance equality of opportunity between persons who share a relevant protected characteristic and persons who do not share it;
- › foster good relations between persons who share a relevant protected characteristic and persons who do not share it.

Notice that these objectives can be seen as having both negative and positive aspects to them (e.g. ‘eliminate discrimination’ as ‘opposed to advance equality’).²

In the context of mental healthcare, where stigmatisation prevents many vulnerable individuals from accessing care and discrimination exacerbates symptoms for already marginalised or minoritised groups, both types of duty are absolutely crucial. However, acting upon positive duties is not without its challenges.

In our workshops, for instance, a key concern that was emphasised was the degree to which the delineation of “good” or “desirable” mental health outcomes, for the purpose of evaluating the impacts of a system, could adequately take into account the subjective nature of mental health and well-being. For instance, while there may be widespread agreement about what constitutes undesirable outcomes (e.g. chronic stress, suicide), the range of positive outcomes for mental health (or well-being) are varied and multitudinous (to echo back to Whitman’s quote that started this report), and the experience and process of recovery can also mean many different things to people³. A failure to account for such issues can introduce a further source of bias into the design of a system (e.g. how the problem it seeks to address is formulated) which in turn could lead to unfair outcomes that only benefit a small group of users with aligned goals.

The pattern that has been developed, through participation of stakeholders and users, attempts to address both negative and positive duties, while also making room for core attributes such as user autonomy. In the context of mental healthcare, this inclusion of autonomy can be problematic in the most severe cases where it is simply not viable (e.g. severe forms of psychosis). However, recall that a pattern is a starting point or scaffold; it is not a checklist of *jointly sufficient claims and supporting evidence*. Therefore, if a particular type of claim is inappropriate due to contextual factors that are determined during project scoping or stakeholder engagement, it can be adjusted as necessary.

Why does this pattern matter?


In recent years, many tools for improving and supporting the trustworthy and responsible development of DMHTs have been proposed. One key advancement is the **Digital Technology Assessment Criteria for Health and Social Care** (DTAC). This form provides guidance on assessing four technical components (in addition to a contextual component):

- 1 Clinical safety
- 2 Data protection
- 3 Technical security
- 4 Interoperability criteria

While the DTAC is intended to supplement existing regulatory guidance, as well as sitting alongside current and developing legislation or compliance duties (e.g. [MHRA medical device registration](#), [Equality Act 2010](#), [NICE's Evidential Standards Framework](#)), there is (at present) nothing in the DTAC about broader ethical issues such as the fair distribution of risk and benefits, or the requirement of explainable outcomes to support informed decision-making. As such, tools such as the DTAC serve a valuable but limited role in the assessment of fair DMHTs.

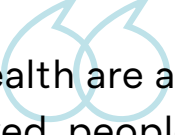
In the last few years, however, many public authorities across the UK have released statements and policies calling for greater health equity. For instance, in October 2020, NHS England released their ['Advancing mental health equalities'](#) strategy, which also fed into a recent consultation on the [Mental health and wellbeing plan](#) by the Department for Health and Social Care.

The second of these publications makes reference to the UK Government's Levelling Up Strategy, which opens with the following statement:



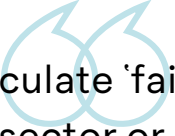
“not everyone shares equally in the UK's success. While talent is spread equally across our country, opportunity is not. Levelling up is a mission to challenge, and change, that unfairness. Levelling up means giving everyone the opportunity to flourish. It means people everywhere living longer and more fulfilling lives, and benefitting from sustained rises in living standards and well-being.”

Beyond people who share protected characteristics as set out in the Equality Act 2010, there are other groups who require specific consideration⁴:



Risks of mental ill-health are also higher for people who are unemployed, people in problem debt, people who have experienced displacement, including refugees and asylum seekers, people who have experienced trauma as the result of violence or abuse, children in care and care leavers, people in contact with the criminal justice system (both victims and offenders), people who sleep rough or are homeless, people with substance misuse or gambling problems, people who live alone, and unpaid carers.

And, finally, a recent publication by the UK's Office for AI has also called for regulators to "embed considerations of fairness into AI", but have further specified this principle as follows:

- 
- › interpret and articulate 'fairness' as relevant to their sector or domain,
 - › decide in which contexts and specific instances fairness is important and relevant (which it may not always be), and
 - › design, implement and enforce appropriate governance requirements for 'fairness' as applicable to the entities that they regulate.

Therefore, there is clear regulatory appetite and industry need for both domain-specific and cross-cutting guidance on how to embed considerations of fairness and equality into the design, development, and deployment of digital technologies.

Fairness Pattern

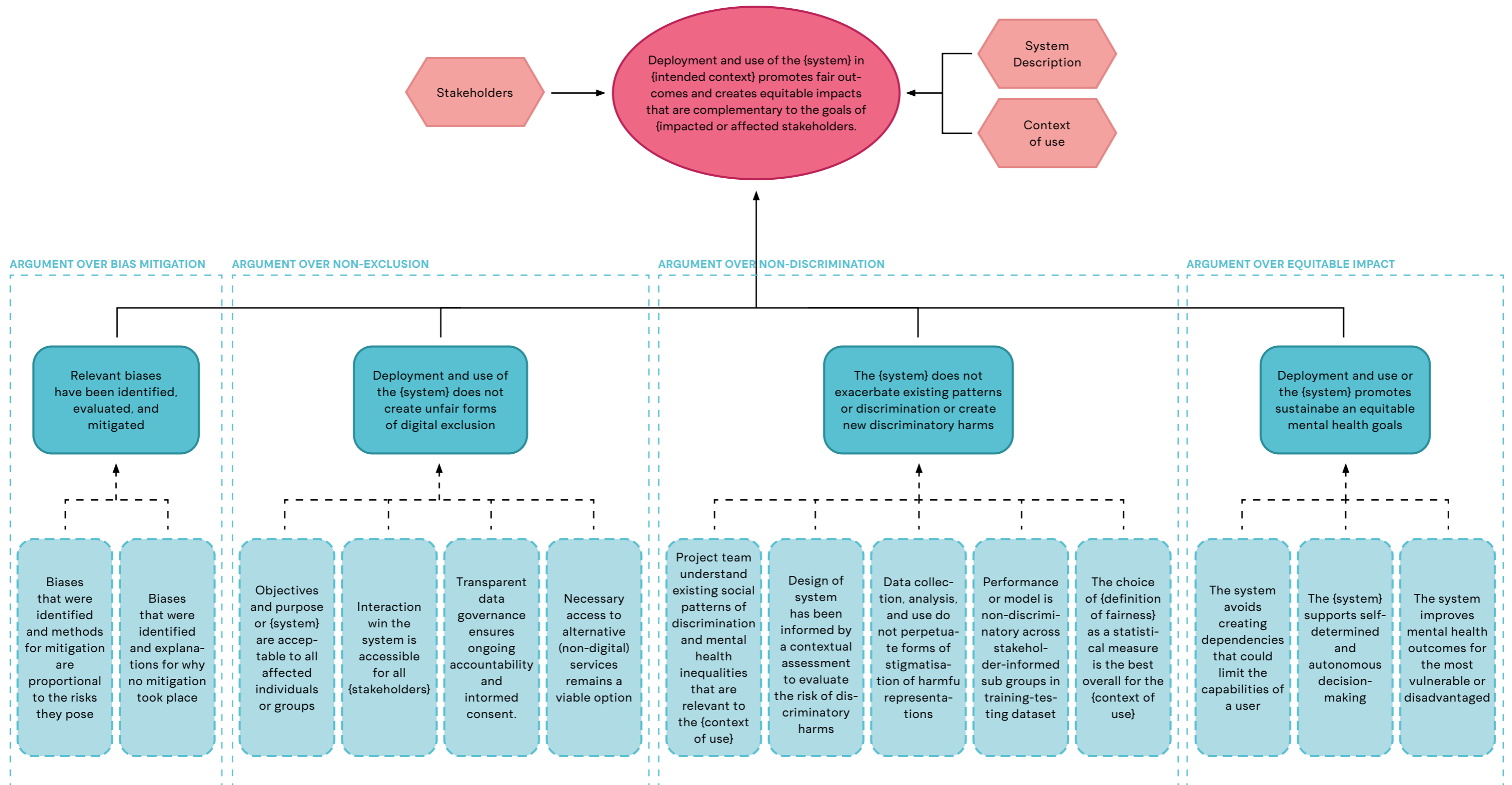


Figure 5.1: A pattern for designing, developing, and deploying fair digital mental health technologies

The goal claim in this argument pattern addresses distributional concepts of fairness (e.g. whether harms and benefits are shared equally), while also acknowledging broader conceptions of social justice (e.g. representational harms to marginalised groups).

Unlike traditional safety cases, which often include System Description and Context of Use placeholders, this pattern also includes a Stakeholder component to emphasise the importance of engagement. Here, the term 'stakeholder' should be treated in as inclusive a manner as possible, and not only the direct users of the technology.

The goal is broken into four higher-level property claims and their respective sub-claims, which we group according to the following core attributes of the Fairness principle as specified and operationalised in the context of digital mental healthcare:

- ➔ **Argument over bias mitigation**
- ➔ **Argument over bias non-exclusion**
- ➔ **Argument over non-discrimination**
- ➔ **Argument over equitable impact**

ARGUMENT OVER BIAS MITIGATION

This argument emphasises identification, evaluation, and mitigation activities. As there are too many biases to incorporate into a single pattern, this argument instead draws attention to transparent and accountable methods that allow stakeholders to determine if the set of biases that have been addressed are sufficient to address their concerns. This argument is supported by additional tools, such as a bias self-assessment tool that outlines social, statistical, and cognitive biases that can affect the lifecycle of a machine learning or AI system project.⁵

ARGUMENT OVER NON-EXCLUSION

This argument sets out another negative duty to consider wider social impacts of digital technologies, and prompts developers to ensure they are not contributing to growing socio-economic inequalities (i.e. the digital divide) by overlooking important social determinants (e.g. education, poverty). In some instances, this may require nothing more than ensuring that UI/UX design choices do not exclude those who have additional accessibility requirements. However, the argument also sets up a duty to consider how a system being developed for use within a public healthcare system, for example, does not create an unsustainable multi-tiered approach, where some users are excluded from accessing a better performing technological service and forced to use comparatively inferior options.

ARGUMENT OVER NON-DISCRIMINATION

This argument addresses the better known obligations, such as ensuring that members of protected groups are not discriminated against. Much of the FairML literature, which has provided useful categorisations of formal criteria (i.e. independence, sufficiency, and separation) and the respective (statistical) notions of individual and group level fairness (e.g. demographic parity, equalised opportunities, counterfactual fairness) are relevant here. And, indeed, a requirement to explain the choice of any statistical measures used during the development of a predictive model (e.g. classifier) is included. Again, there are useful tools and taxonomies that offer further guidance on these decisions. However, our pattern also urges project teams to reflect upon other patterns of discrimination, marginalisation, and minoritisation, which can exacerbate mental health issues, but which fall beyond protected characteristics that lie outside the scope of the Equality Act 2010 (e.g. housing, employment, social support). Doing so will typically require engagement of domain experts where such expertise does not exist within teams.

ARGUMENT OVER EQUITABLE IMPACT

Finally, this argument references positive duties that matter in the context of digital mental healthcare specifically. Key to this is the consideration of ethical values such as autonomy and self-determination, and the prioritarian weighting that was mentioned previously. However, there is also an aspect of our Sustainability principle that trickles into this argument. This is important in the context of digital mental healthcare because of associated risks that a) positive effects diminish over time (e.g. behavioural nudges or habit forming techniques that become ineffective over time)⁶ and b) negative impacts worsen and compound (e.g. prolonged use of social media worsening anxiety or depression)⁷. Studies have already criticised the evidence base of mental health apps and services⁸, especially for a lack of reliable longitudinal evidence, so drawing attention to sustainable impacts and setting up requirements for continuous monitoring is vital to maintain trust and also ensure that specific users are not locked in to services or technologies that degrade in quality or efficacy over time (e.g. apps that start by offering free services to leverage network effects only to force subscriptions at a later date).

Explainability



Much like fairness, explainability has become a popular and thriving area of research (e.g. so-called XAI). And, also like fairness, it is a normative goal that encompasses a range of significant concerns and salient ethical values.

The most obvious conceptual distinction is between interpretability as a core component of explainability. Whereas the former is often treated as the ability for developers or users to understand the inner workings of algorithms (or inability in the case of complex, non-linear techniques), the latter refers to an interpersonal ability to communicate knowledge in a manner that is accessible to those who may be asking questions about a system (e.g. patients asking for explanations from a clinician). Although statistical techniques help significantly in the case of the former, they are more limited in their ability to support the latter where a more diverse range of users are likely.

Why does this pattern matter?

Explainable AI has received a lot of attention over the last several years.⁹ Computer scientists have developed new tools and methods to improve the interpretability of otherwise opaque algorithms, such as neural networks.¹⁰ Researchers in psychology and human-computer interaction have explored how different components of the user experience can help support more intentional interactions with intelligent software agents.¹¹ And, regulators, auditors, and journalists have investigated how to make systems more transparent to support objectives related to accountability and informed decision-making.¹²

Much of this attention arises from the recognition that data-driven technologies have the potential to automate decision-making to varying degrees and, therefore, affect key ethical


values and principles such as autonomy, accountability, responsibility, and informed consent. On the one hand, decision support systems can offer recommendations to users but are not responsible for enacting a decision directly. And, on the other hand, you have *fully automated-decision making systems*, which once set up require no human involvement.

This distinction is admittedly coarse grained, but it will suffice for our purposes because it helps identify two illustrative cases where explainability matters. In the former case, although a human user is responsible for the decision, their judgement may be influenced and biased by the decision support system, potentially in ways that are problematic (e.g. leading to differential treatment for certain groups of users). In the latter case, no human is involved, but because the automated systems cannot be held morally or legally accountable for their decisions, if something goes wrong, a human will need to be able to identify the reason why the problem occurred and perhaps communicate this to other affected stakeholders.

In both of the above cases, extracting a valid and accurate explanation is necessary to enable post hoc forms of accountability or transparency. But prioritising 'explainability' from the start of a project also allows project teams to have better oversight of what their systems do and why, leading to more responsible forms of project governance. And, at the other end of the lifecycle, clear and accessible explanations can help ensure users and affected stakeholders are better *informed* and empowered to make *autonomous decisions* regarding their interactions with DMHTs. Therefore, having an argument pattern for 'explainability' helps capture many of the key considerations that were raised during our workshops.

While the themes and values expressed in the following pattern are based primarily on the engagement with stakeholders, we have also drawn upon two other documents. First, we have drawn from prior regulatory guidance that we co-designed with the Information Commissioner's Office. This guide, titled 'Explaining Decisions Made with AI', details best practices for explainable AI in domain-general settings, and was also informed by stakeholder engagement. The regulatory ecosystem around explainability is less developer than fairness and equality, but as this report acknowledges there are still legislative and regulatory considerations that organisations need to consider, such as the wide-range of rights established by the General Data Protection Regulation and implemented in the UK's Data Protection Act 2018, such as the need to uphold individuals rights to be informed or to object to automated decisions.¹³

Second, we have incorporated some elements of an existing pattern for *interpretable* machine learning¹⁴, which is motivated by a similar need for addressing a range of questions and concerns, such as the following:



“who needs to understand the system, what they need to understand, what types of interpretations are appropriate, and when do these interpretations need to be provided” (Ward and Habli, 2020).

Explainability Pattern

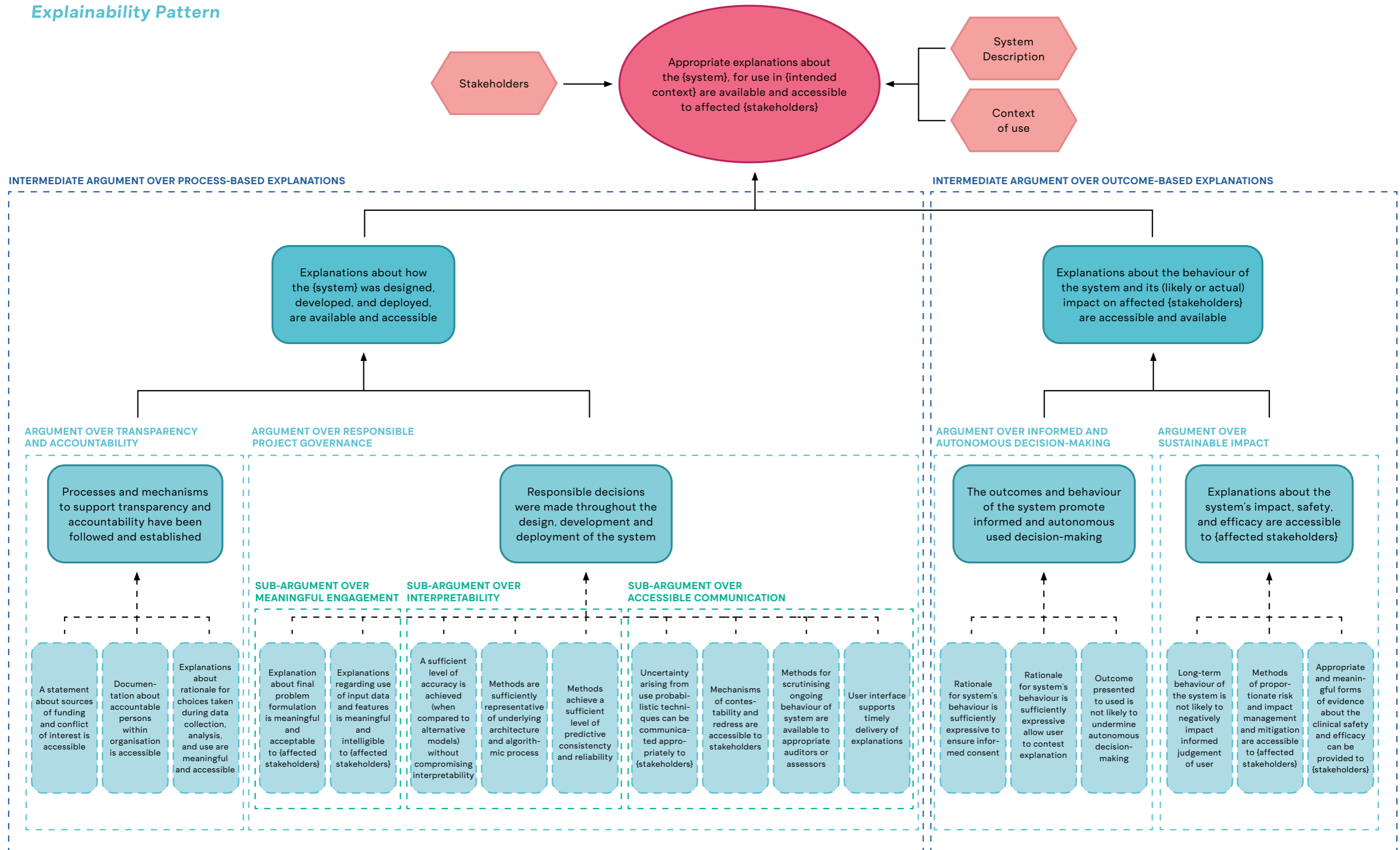


Figure 5.2: A pattern for designing, developing, and deploying explainable digital mental health technologies

In previous guidance¹⁵, we have distinguished between two sub-categories of explanations:

-
- 1 Process-based explanations of AI systems are about demonstrating that you have followed good governance processes and best practices throughout your design and use.
 - 2 Outcome-based explanations of AI systems are about clarifying the results of a specific decision. They involve explaining the reasoning behind a particular algorithmically-generated outcome in plain, easily understandable, and everyday language.
-

These categories are reflected in our pattern, where they form 'intermediate arguments' that help refine the goal claim and also serve as scaffolding for the main arguments.

As with the fairness pattern, placeholders for System Description and Context of Use and Stakeholder are also included.

The intermediate arguments are then broken into four higher-level property claims and their respective sub-claims, which we group according to the following core attributes of the Explainability principle as specified and operationalised in the context of digital mental healthcare:

- ➔ **Argument over transparency and accountability**
- ➔ **Argument over responsible project governance**
- ➔ **Argument over informed and autonomous decision-making**
- ➔ **Argument over sustainable impact**

ARGUMENT OVER TRANSPARENCY AND ACCOUNTABILITY

This argument addresses the processes and mechanisms that have been undertaken throughout the project lifecycle to establish sufficient forms of transparency and accountability. This includes documentation relevant to the identification of responsible project members, as well as choices made about data (e.g. why certain data types were included or excluded). Importantly, this argument also recommends the inclusion of a statement about sources of funding and conflicts of interest, which was an important matter for trustworthiness that arose during our engagement with participants with lived experience of DMHTs.

This argument is more comprehensive than the others, and so is further split into three sub-arguments:

-
- 1** Sub-argument over meaningful engagement: here, meaningful engagement can be seen to include participation in decisions about the formulation of the problem that a DMHT is expected to address, as well as issues of data usage—both of which affect later stages of the project lifecycle and the final behaviour of the deployed system.

 - 2** Sub-argument over interpretability: sufficient levels of accuracy and the potential trade-off with interpretability can require high levels of technical and data literacy. Therefore, this argument focuses on the requisite information that is needed to support explainability (recall earlier distinction between interpretability and explainability).

 - 3** Sub-argument over accessible communication: the previous sub-argument feeds into this sub-argument, which focuses on how ‘accessible’ forms of communication will be achieved, and the challenges of communicating probabilistic information. Ultimately, this sub-argument will depend on decisions reached and evidence obtained through consultation with intended users, as well as on the basis of knowledge about any time-constraints presented by the context of use (e.g. urgency in high-risk care environments).

ARGUMENT OVER INFORMED AND AUTONOMOUS DECISION-MAKING

The core attribute motivating this argument is shared with our fairness pattern. Here, the argument emphasises the importance of explanations that refer to the observed behaviours or outcomes of the system. For instance, one of the claims is intended to ensure that explanations are “sufficiently expressive”, without overwhelming the user with unnecessary or overly-complex information. This will depend on the intended user and context of use. However, to supplement this claim, emphasis is also placed on the ability for user to challenge outcomes, rather than just having them explained without an option to contest.

ARGUMENT OVER SUSTAINABLE IMPACT

The final argument also follows the theme of the fairness pattern, but rather than addressing equitable impact, it focuses on sustainable impact.¹⁶ This is important because explanations are sometimes used as a means to justify why a specific norm was transgressed (e.g. why you were late). However, over time, if the same explanation is provided without a change to the offending behaviour, the explanation loses its validity. A similar risk is present in the automated delivery of explanations by algorithmic systems. For example, if an AI chatbot continues to offer the same inaccurate and irrelevant explanations, it is likely to lose the trust of a user. Therefore, assessments about the impact of explainable AI need to account for longer-term dynamics to ensure that the relevant systems are sustainable over time.

The inclusion of clinical safety and efficacy should not suggest that these goals are not significant in their own right. In fact, we would advocate for a separate assurance case (and corresponding pattern) on these goals specifically. Instead, reference is simply made to the need to ensure that some form of explanation is provided to stakeholders.

Evidential Considerations

Neither of the patterns above include prescriptions about specific evidential artefacts that could be used to ground the assurance case. There are two reasons for this intentional omission:

-
- 1 Prescribing specific forms of evidence is too difficult outside of highly constrained contexts where there are clear details about a) the intended use context, b) the type of ML/AI technique being used, and c) the intended users and target audience.
-
- 2 Developers and regulators should be free to determine the appropriate forms of evidence, based on developing best practices and standards, many of which do not exist at present.

However, there are a couple of general remarks that can be made, as well as some suggestions for further resources.

First, as we have argued elsewhere¹⁷, the generation, evaluation, and selection of evidence can be guided by the following considerations:

-
- 1 Is the evidential artefact/claim relevant to the parent claim?
-
- 2 Is the evidential artefact/claim (or set of artefacts/claims) sufficient to justify the parent claim?
-
- 3 Is there sufficient probative value in the overall assurance case to justify the top-level normative goal?

Outside of the regulatory considerations already mentioned above, there have been several developments in recent years to help organisations address these considerations. Some examples (among many) include:

- **Model Cards for Model Reporting**: templates for model documentation, which include ethical considerations alongside statistical information to support reuse.
- **Responsible AI Licensing**: licenses that help developers restrict the use of their AI technology in order to prevent irresponsible and harmful applications.

- **Data Hazards**: a set of labels that enable project members to make decisions about the risks of data-driven technologies using a shared vocabulary
- **Assurance of Machine Learning in Autonomous Systems (AMLAS)**: a methodology for assuring the safety of ML systems, with systematic means for evaluating processes such as model testing or verification.
- **Algorithmic Transparency Standard**: a template for organisations to use when choosing to publish information about how they are using algorithmic systems to aid decision-making.

Typically, these tools, methods, or templates exist to help organisations and project teams address a specific challenge (e.g. licensing, model documentation). The framework and methodology we have presented in this report is designed to work alongside these tools, but it also goes further by helping teams organise them according to a particular goal or objective (e.g. fairness). As such, our framework and methodology is broader in scope and offers a systematic means for choosing when to use specific tools throughout a project's lifecycle and how to bring the documented evidence together to create a trustworthy and justifiable assurance case.

Conclusion



Co-Creating a Culture of Trust

Uncertainty breeds distrust.

When the consequences of potential actions or interventions are hard to identify or evaluate, inaction or inertia can follow. And, such uncertainty and deliberative inertia is only exacerbated in the context of mental health (e.g. challenges that are faced by those with depression or anxiety disorders, such as catastrophising, when attempting to evaluate actions).

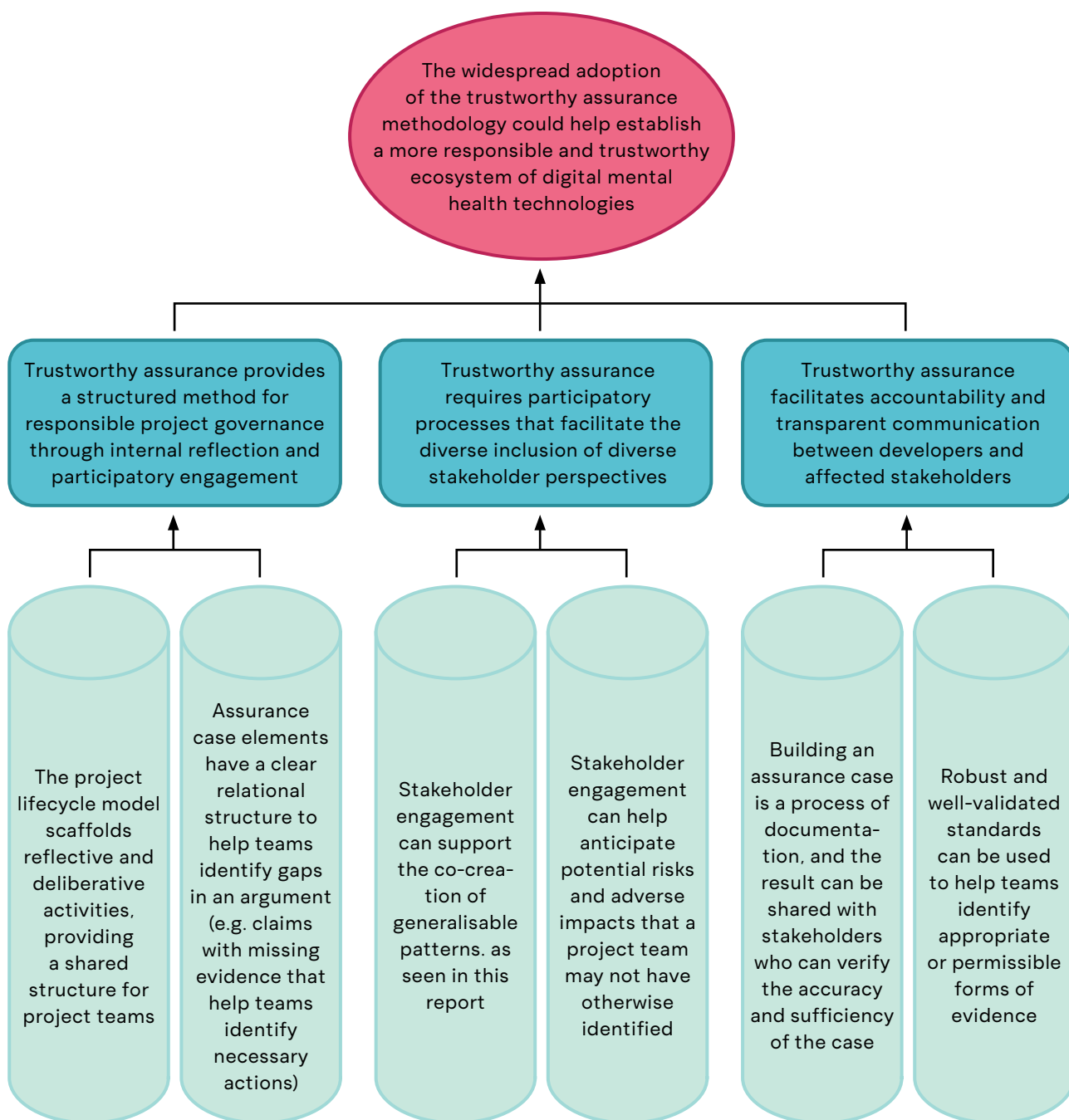
As we have seen over the course of this report, vague privacy policies, poorly-specified objectives, pervasive and invasive data extraction, and dubious claims about user safety and clinical efficiency of services, are all sources of uncertainty.

This report has laid out a methodological proposal for how we can begin to address the current culture of distrust that casts a shadow over the digital mental healthcare landscape. While we believe that the methodology and recommendations will support this goal, and have presented tentative evidence to justify this belief, the development of a trustworthy ecosystem of digital mental healthcare requires a more collaborative effort.

Therefore, in addition to our methodological proposal of trustworthy assurance and the initial argument patterns, we have also made a series of supporting recommendations (summarised in the [Executive Summary](#)). But what comes next?

Next Steps

In addition to the individual recommendations presented in the previous chapters, this report can also be viewed as a general recommendation itself—one that calls for the more widespread adoption of the trustworthy assurance methodology. We can even formulate this recommendation as an argument in a pseudo assurance case:



Ultimately, the veracity of the goal claim depends on how and whether the methodology is adopted. This report is not a user guide, however, so more work needs to be done to ensure the adoption by organisations is made as straightforward as possible (e.g. alignment with existing regulation and current practices, including complementary quality assurance procedures).

A key next step will be to develop user guidance that can help with this objective. This has already commenced, and we currently have a) a full-length article that goes into further detail about the methodology¹, in relation to domain-general ethical principles, and b) a prototype platform that can enable the production of assurance cases². However, these proposals have hitherto not been connected directly with the more formal work undertaken by those working in safety assurance. This is primarily because we endeavoured to make the trustworthy assurance methodology simple and accessible for the purpose of our stakeholder engagement. But, the adoption of the methodology by developers and engineers would likely benefit from closer integration with standardisation efforts, such as the Goal Structuring Notation (GSN) and the [GSN Standard Working Group](#).

These efforts will need to remain receptive to ongoing developments in this domain, whether technological (e.g. development of new computational techniques or devices), legislative (e.g. reforms to UK legislation such as the Draft Mental Health Bill, 2022), regulatory (e.g. report on the [Public Sector Equality Duty](#) by the Equality and Human Rights Commission, the formation of the [Multi Agency Advice Service](#); proposal of the [Digital Technology Assessment Criteria \(DTAC\)](#), or societal (e.g. changing perceptions or attitudes of users towards mental health and well-being services, including data-driven technologies).

We hope that the proposal and recommendations set out in this report offers some clarity, structure, and positive direction to help navigate this complex (and multitudinous) space. And, more importantly, we hope that it can (indirectly) improve mental health services for those who need them.

Further Resources

To keep the length of this version of our report to a minimum, we have included two (optional) appendices in the online version only:

- › **Appendix 1** covers our project's methodology.
- › **Appendix 2** provides illustrative examples of DMHTs.

They can be accessed here: <https://alan-turing-institute.github.io/trustworthy-assurance/dmh-report/about/>



Endnotes

Foreword

- 1 <https://www.businessofapps.com/data/>

Executive Summary

- 1 Work is underway to develop an user guide, and this will be added to the online version of our report when ready. The guide will also include instructions on how to use our tool for producing assurance cases.
- 2 The following documents provide a more critical examination for those interested: ([Sujan and Habli, 2021](#)); ([Burr and Leslie, 2022](#)).
- 3 Christopher Burr, J. Morley, M. Taddeo, and L. Floridi. Digital Psychiatry: Risks and Opportunities for Public Health and Wellbeing. IEEE Transactions on Technology and Society, 1(1):21–33, March 2020. [doi:10.1109/TTS.2020.2977059](https://doi.org/10.1109/TTS.2020.2977059).

Introduction

- 1 See for example, the work programme being conducted by the Medicines & Healthcare Regulatory Agency (MHRA) on [Software and AI as a Medical Device](#), the [Evidence standards framework](#) (ESF) for digital health technologies from the National Institute for Health and Care Excellence (NICE), and the proposed work from the [Multi-agency advisory service](#) (MAAS) for artificial intelligence (AI) and data-driven technologies, which is being funded by the [NHS AI Lab](#).
- 2 For the purpose of this report we use the term ‘trust’ to refer to those characteristics of a person’s beliefs or attitudes that are directed towards an object, person, or proposition (among other things), whereas ‘trustworthiness’ refers to the perceived property or attribute which an individual uses to determine whether to place trust (e.g. whether to trust a news article based on its quoted sources).

- 3 Burr, C., J. Morley, M. Taddeo, & L. Floridi. (2020). Digital Psychiatry: Risks and Opportunities for Public Health and Wellbeing. IEEE Transactions on Technology and Society, 1(1), 21–33. <https://doi.org/10.1109/TTS.2020.2977059>
- 4 We explored this point more fully in a separate article: Burr, C. (2022). Charities are contributing to growing mistrust of mental-health text support—Here’s why. The Conversation. Retrieved 29 July 2022, from <http://theconversation.com/charities-are-contributing-to-growing-mistrust-of-mental-health-text-support-heres-why-179056>
- 5 Nuffield Council on Bioethics (2022) The role of technology in mental healthcare. <https://www.nuffieldbioethics.org/assets/pdfs/The-role-of-technology-in-mental-healthcare.pdf>
- 6 Torous et al. (2016) New Tools for New Research in Psychiatry: A Scalable and Customizable Platform to Empower Data Driven Smartphone Research. JMIR Ment Health. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4873624/>
- 7 For instance, if we draw the class as comprising digital technologies for “health and well-being” we will capture technologies as diverse as ML algorithms used to identify associations between genetic factors and mental health outcomes, to mobile apps that use natural language processing techniques to help users better understand their feelings through a smart diary.
- 8 <https://www.england.nhs.uk/mental-health/cyp/>
- 9 <https://digital.nhs.uk/data-and-information/publications/statistical/mental-health-of-children-and-young-people-in-england/2021-follow-up-to-the-2017-survey>
- 10 <https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/bulletins/suicidesintheunitedkingdom/2020registrations>

Endnotes

- 11 <https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/articles/coronavirusanddepressioninadultsgreatbritain/januarytomarch2021>
- 12 See references and discussion on 'universal design' in Burr, C., Taddeo, M., & Floridi, L. (2020). The Ethics of Digital Well-Being: A Thematic Review. *Science and Engineering Ethics*. <https://doi.org/10.1007/s11948-020-00175-8>
- 13 For a timeline that conveys a shocking pattern of behaviour at Facebook, which is hard to treat as anything other than a flagrant disregard for user's data privacy, see [Facebook data privacy scandal: A cheat sheet](#).
- 14 It is important to note that private organisations are still bound by the legal duties of non-discrimination, even where they fall outside the wider scope of the public sector equality duty. For further advice and guidance on these topics, see [this recent publication](#) by the Equality and Human Rights Commission on Artificial Intelligence in Public Services.
- 15 The [Equality Act 2010](#) sets out the following protected characteristics:
- Age
 - Disability
 - Gender reassignment
 - Marriage and civil partnership
 - Pregnancy and maternity
 - Race
 - Religion or belief
 - Sex
 - Sexual orientation
- They are protected in the sense that the law is designed to protect individuals from unfair treatment or discrimination on the basis of these characteristics.
- 16 Our thanks to a reviewer for bringing the following example to our attention of a local council who unanimously voted to make 'care experience' a protected characteristic within its constituency: <https://www.cypnow.co.uk/news/article/cumberland-council-votes-to-make-care-experience-a-protected-characteristic>.
- 17 For instance, Article 8 of the EU Charter of Fundamental Rights (On the Protection of personal data) states, "1. Everyone has the right to the protection of personal data concerning him or her. 2. Such data must be processed fairly for specified purposes and on the basis of the consent of the person concerned or some other legitimate basis laid down by law. Everyone has the right of access to data which has been collected concerning him or her, and the right to have it rectified. 3. Compliance with these rules shall be subject to control by an independent authority." Provisions 1 and 2 help to delineate the remit of an individual's right, whereas provision 3 establishes a corresponding duty on member states that helps ensure the aforementioned rights are guaranteed and protected.
- 18 Dark patterns are design elements of a user interface that have been intentionally chosen to manipulate or deceive users into taking actions or making sub-conscious choices, which they would be unlikely to do when conscious of the outcome (e.g. purchasing a more expensive product, agreeing to invasive privacy policies).
- 19 To emphasise the consistency of this recommendation with the conclusions from the Nuffield Council report we should also acknowledge that this 'digital ecosystem' must be complementary with, supportive of, and enhance more traditional healthcare services (qua conclusion 2).
- 20 See an older, but still important report from [Health Foundation](#) as well as a more recent proposal: Habli et al. (2020). Enhancing COVID-19 decision making by creating an assurance case for epidemiological models. *BMJ Health & Care Informatics*, 27(3), e100165. <https://doi.org/10.1136/bmjhci-2020-100165>

Endnotes

- 21 The most recent and comprehensive account of this framework can be found in the following proposal to the Council of Europe: Leslie, D., Burr, C., Aitken, M., Katell, M., Briggs, M., & Rincon, C. (2022). Human rights, democracy, and the rule of law assurance framework for AI systems: A proposal. <https://rm.coe.int/huderaf-coe-final-1-2752-6741-5300-v-1/1680a3f688>.
- 22 Christopher Burr and David Leslie. Ethical assurance: a practical approach to the responsible design, development, and deployment of data-driven technologies. *AI and Ethics*, June 2022. URL: <https://link.springer.com/10.1007/s43681-022-00178-0> (visited on 2022-08-01), [doi:10.1007/s43681-022-00178-0](https://doi.org/10.1007/s43681-022-00178-0)

Chapter 2

- 1 Rather, it “takes a village to raise an algorithm” as noted in this [blog post](#). Thanks to one of our reviewers for bringing this post to our attention.
- 2 To be clear, we are referring here to role and task responsibilities first and foremost, which overlap with but are conceptually separate from moral responsibility.
- 3 Updating can also occur in cases where no adaptive behaviour is present for reasons such as performance improvements or bug fixes.
- 4 See Part 1 of the [GSN Standard, Version 3](#) for a complete overview of the notation.
- 5 Another alternative to GSN is the [Claims, Arguments and Evidence](#) notation developed by Adelaar.
- 6 For example, guidance such as ISO/IEC/IEEE 16085, ‘Systems and software engineering - Life cycle management - Risk Management’
- 7 The AI Standards Hub, for example, serves an observatory of relevant standards for AI technologies (see <https://aistandardshub.org>).
- 8 For an example of this type of technology, see Lucas et al. (2017). Reporting Mental Health Symptoms: Breaking Down Barriers to Care with Virtual Human Interviewers. *Frontiers in Robotics and AI*. <https://www.frontiersin.org/articles/10.3389/frobt.2017.00051>
- 9 See Kirsch, A. (2017). Explain to whom? Putting the User in the Center of Explainable AI. Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML 2017 Co-Located with 16th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2017). <https://hal.archives-ouvertes.fr/hal-01845135>
- 10 Recall, that evidential claims are required to link property claims to their supporting evidential artefacts, and evidential claims, therefore, can also serve as reasons.
- 11 Ward, F. R., & Habli, I. (2020). An Assurance Case Pattern for the Interpretability of Machine Learning in Safety-Critical Systems. https://doi.org/10.1007/978-3-030-55583-2_30
- 12 Nickel, J. W. (2007) *Making Sense of Human Rights* (2nd Edition). Blackwell Publishing.
- 13 The most recent and comprehensive account of this framework can be found in the following proposal to the Council of Europe: Leslie, D., Burr, C., Aitken, M., Katell, M., Briggs, M., & Rincon, C. (2022). Human rights, democracy, and the rule of law assurance framework for AI systems: A proposal. <https://rm.coe.int/huderaf-coe-final-1-2752-6741-5300-v-1/1680a3f688>.
- 14 For a contrasting approach, see Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*. <https://doi.org/10.1162/99608f92.8c-d550d1>

Endnotes

- 15 This process is (loosely) derived from the idea of 'reflective equilibrium', made famous by the political philosopher John Rawls. In short, the phrase 'reflective equilibrium' refers to a state of coherence among moral beliefs and attitudes, which emerges over time as a result of public deliberation and consensus building activities that focus on the relevant moral beliefs and attitudes (e.g. notions of 'justice').
- 16 Richard Hawkins, Colin Paterson, Chiara Picardi, Yan Jia, Radu Calinescu, and Ibrahim Habli. Guidance on the Assurance of Machine Learning in Autonomous Systems (AMLAS). Technical Report, University of York, Assuring Autonomy International Programme, March 2021. URL: <https://www.york.ac.uk/media/assuring-autonomy/documents/AMLASv1.1.pdf> (visited on 2022-09-08).
- 17 The Assurance Case Working Group. GSN Community Standard Version 3. Technical Report, The Assurance Case Working Group, May 2021. URL: <https://scsc.uk/r141C:1?t=1>.

Chapter 3

- 1 Shackle, S. (2019). 'The way universities are run is making us ill: inside the student mental health crisis. The Guardian. <https://www.theguardian.com/society/2019/sep/27/anxiety-mental-breakdowns-depression-uk-students>
- 2 Kotouza, D., Callard, F., Garnett, P., & Rocha, L. (2022). Mapping mental health and the UK university sector: Networks, markets, data. *Critical Social. Policy*, 42(3), pp.356-387 <https://journals.sagepub.com/doi/full/10.1177/02610183211024820>
- 3 Williams, T. (2022, May 9). UK student mental health crisis 'may be worse than thought'. *Times Higher Education*. <https://www.timeshighereducation.com/news/uk-student-mental-health-crisis-may-be-worse-thought>; <https://www.if.org.uk/2021/03/11/new-ons-data-show-student-mental-health-crisis/>

- 4 Office for Students (2019) Mental Health: Are all students being properly supported? <https://www.officeforstudents.org.uk/media/b3e6669e-5337-4caa-9553-049b3e8e7803/insight-brief-mental-health-are-all-students-being-properly-supported.pdf>; Duffy, A., Saunders, K.E.A, Malhi, G.S., Patten S., Cipriani, A., McNevin, S. H., MacDonald, E., & Geddes, J. (2019) Mental health care for students: A way forward? *The Lancet Psychiatry*, 6(11). [https://www.thelancet.com/journals/lanpsy/article/PIIS2215-0366\(19\)30275-5/fulltext](https://www.thelancet.com/journals/lanpsy/article/PIIS2215-0366(19)30275-5/fulltext)
- 5 Gunnell, D., Kidger, J., & Elvidge, H. (2018). Adolescent mental health in crisis. *BMJ* 361. <https://www.bmj.com/content/361/bmj.k2608.long>
- 6 National Union of Students Northern Ireland. (2017). NUS-USI Student Wellbeing Research Report, https://nusdigital.s3-eu-west-1.amazonaws.com/document/documents/33436/59301ace47d6320274509b83e-1bea53e/NUSUSI_Student_Wellbeing_Research_Report.pdf
- 7 Thorley, C., (2017). Not by degrees: Improving student mental health in the UK's universities, Universities UK (2020). Stepchange: Mentally healthy universities., Hughes, G., & Spanner, L. (2019). The University Mental Health Charter.
- 8 Cate, N. (2021, August 16). The role of digital mental health support tools and the importance of the student co-productin model in supporting their development. Office for Students. <https://www.officeforstudents.org.uk/advice-and-guidance/student-wellbeing-and-protection/student-mental-health/the-role-of-digital-mental-health-support-tools-and-the-importance-of-the-student-co-production-model-in-supporting-their-development/>
- 9 NHS X (2019). Artificial Intelligence: How to get it right? https://www.nhs.uk/media/documents/NHSX_AI_report.pdf

Endnotes

- 10 Cate, N. (2021, August 16). The role of digital mental health support tools and the importance of the student co-productin model in supporting their development. Office for Students. <https://www.officeforstudents.org.uk/advice-and-guidance/student-wellbeing-and-protection/student-mental-health/the-role-of-digital-mental-health-support-tools-and-the-importance-of-the-student-co-production-model-in-supporting-their-development/>
- 11 These figures are based upon desk research by The Alan Turing Institute team, looking at universities' public webpages. In these cases, a university's recommendation of a service does not indicate that any formal relationship exists with the service in question. At times these recommendations for self-help apps are accompanied by a disclaimer. For example, the University of Durham notes that "the counselling service is happy to signpost this information/the availability of free access to these resources, in the hope they might prove useful to students and staff. However, please be aware that the Counselling Service does not have any relationship or affiliation with any of the providers, nor does it support, or endorse in any way the external information, advice, other services and/or resources that can be accessed via the links below (<https://www.dur.ac.uk/counselling.service/self-help/>). In other instances ,they are simply offered as resources for students to browse at their own discretion.
- 12 Kotouza, D., Callard, F., Garnett, P., & Rocha, L. (2022). Mapping mental health and the UK university sector: Networks, markets, data. *Critical Social. Policy*, 42(3), pp.356-387 <https://journals.sagepub.com/doi/full/10.1177/02610183211024820>
- 13 Bennett, N. (2021, February 15) Rebuilding post-16 education around mental fitness. Fika community. <https://www.fika.community/insight/rebuilding-post-16-education-around-mental-fitness>
- 14 <https://student.kooth.com>
- 15 Nuffield Council on Bioethics. (2022). The role of technology in mental healthcare. <https://www.nuffieldbioethics.org/assets/pdfs/The-role-of-technology-in-mental-healthcare.pdf>
- 16 <https://www.amosshe.org.uk/futures-duty-of-care-2015>
- 17 AMOSSHE. (2015). Where's the line? How far should universities go in providing duty of care for their students. <https://www.amosshe.org.uk/futures-duty-of-care-2015>
- 18 There is a huge and varied literature on 'algorithmic bias', and so pinning down a single definition of the term is challenging. In other works we have instead focused on exploring three types of bias—social, statistical, and cognitive, which can impact the design, development, and deployment of ML and AI technologies. See [here](#) for further information.

Chapter 4

- 1 There is, of course, a trade-off here between the narrower and wider perspectives, which we discuss in [Appendix 1](#).
- 2 This is not to say that clinical efficacy is not a key component of ethical decision-making. For instance, the bioethical principle of beneficence clearly requires sufficient levels of clinical efficacy.

Endnotes

- 3 These objectives were as follows: 1) To explore whether and how the methodology of argument-based assurance could be extended to address ethical issues in the context of digital mental healthcare. 2) To evaluate how an extension of the methodology could support stakeholder co-design and engagement, in order to build a more trustworthy and responsible ecosystem of digital mental healthcare. 3) To lay the theoretical and practical foundations for scaling the ethical assurance methodology to new domains, while integrating wider regulatory guidance (e.g., technical standards).
- 4 The attentive reader will see significant overlap between the concepts that are mapped onto the principles, and subsequently the principles themselves (e.g. explainability and accountability). Because the principles were not designed to be mutually exclusive and collectively exhaustive, this overlap is to be expected.
- 5 Beauchamp, T. L., & Childress, J. F. (2013). Principles of biomedical ethics (7th ed.). Oxford University Press.
- 6 Leslie, D. et al. (2022). Human rights, democracy, and the rule of law assurance framework for AI systems: A proposal. Accessed: <https://rm.coe.int/huderaf-coe-final-1-2752-6741-5300-v-1/1680a3f688>
- 7 Our case studies included a section on 'Affected individuals, groups, and other stakeholders'. For this case study, 'psychiatrists' were included (see [case study 3](#).)
- 8 See our report on developing '[Common Regulatory Capacity for AI](#)' for more on these topics.
- 9 For transparency, the distribution of responses by stakeholder group is as follows: Strongly agree: policy-maker x2, researcher x1, developer x1; Agree: policy-maker x1, researcher x4, developer x2; Undecided: policy-maker x2
- 10 Again, these results should be treated with caution due to the small sample size. Strongly agree: policy-maker x1, researcher x1, developer x2; Agree: policy-maker x2, researcher x4, developer x4; Undecided: policy-maker x2, developer x1
- 11 See Equality and Human Rights Commission. (2014, August 1). Technical Guidance on the Public Sector Equality Duty: England | Equality and Human Rights Commission. Technical Guidance on the Public Sector Equality Duty: England. <https://www.equalityhumanrights.com/en/publication-download/technical-guidance-public-sector-equality-duty-england>
- 12 O'Neill, O. (2002). A Question of Trust. Cambridge University Press, pp. 77–78.
- 13 Charlton, J. I. (1998). Nothing about us without us: Disability Oppression and Empowerment. University of California Press.
- 14 This theme is built into an argument pattern in the next chapter. The pattern urges reflection upon and consideration of deep patterns of discrimination, marginalisation, and minoritisation, which can exacerbate mental health issues, but which fall beyond protected characteristics that lie outside the scope of the Equality Act 2010 (e.g. poverty, housing, employment).
- 15 On these points it is worth noting that the [Mental Health Act](#), which sets out the legislation that is used to determine when it is appropriate to place restrictions on people, has been subject to proposed reform in recent months. Readers may find the "new guiding principles" of particular interest in the context of this report (see here).
- 16 For instance, the goal of increasing health equality is incorporated into a recent discussion paper from the Department of Health and Social Care's [Mental health and wellbeing plan](#), the '[Advancing mental health equalities](#)' strategy from NHS England, the Welsh Government's '[Together for Mental Health](#)' delivery plan, and an ongoing consideration for the Scottish Government following the publication of an equality impact assessment of their mental health strategy back in 2017

Endnotes

Chapter 5

- 1 In a previous article we also explore several considerations about the evidence generation and selection process, including whether evidential artefacts are permissible, sufficient, and relevant. See Burr, C., & Leslie, D. (2022). Ethical assurance: A practical approach to the responsible design, development, and deployment of data-driven technologies. *AI and Ethics*. <https://doi.org/10.1007/s43681-022-00178-0>
- 2 We are not suggesting that this is the sole correct way of interpreting the public sector equality duty. Others may view the associated duties as entirely positive if they are viewed from a human rights lens that treats actively protecting people from risks of harm that are known about, or should have been known about, as a positive duty. The status of this duty will likely depend on which party it falls on, and how they are expected to discharge the duty.
- 3 For instance, a patient experiencing depression may be fully informed by their psychiatrist about their mental health and the options available to them in terms of recovery, but nevertheless, autonomously decide to forego any treatment because their condition may be an important part of their self-identity. This acknowledgment is part of the recovery approach, which views recovery as “a deeply personal, unique process of changing one’s attitudes, values, feelings, goals, skills, and/or roles. It is a way of living a satisfying, hopeful, and contributing life even with limitations caused by illness. Recovery involves the development of new meaning and purpose in one’s life as one grows beyond the catastrophic effects of mental illness.” W. A. Anthony, “Recovery from mental illness: The guiding vision of the mental health service system in the 1990s,” *Psychosoc. Rehabil. J.*, vol. 16, no. 4, p. 527, 1993, doi: 10.1037/h0095655.
- 4 This is further problematised in the context of ML and AI, where novel forms of discrimination, perhaps as a result of so-called “affinity profiling”. See Wachter, S. (2021). Affinity Profiling and Discrimination by Association in Online Behavioural Advertising. *Berkeley Technology Law Journal*, 35(2).
- 5 See our course on responsible research and innovation for more details about social, statistical, and cognitive biases: <https://alan-turing-institute.github.io/turing-commons/rri/>
- 6 See Maier et al. (2022) No evidence for nudging after adjusting for publication bias. *PNAS*. <https://doi.org/10.1073/pnas.2200300119>
- 7 For a review published prior to the onset of the COVID-19 pandemic see Keles (2019) A systematic review: the influence of social media on depression, anxiety and psychological distress in adolescents. *Adolescence and Youth*. <https://doi.org/10.1080/02673843.2019.1590851>. For a range of studies that focus on the impacts of COVID-19 on mental health, including several that explore social media, see the [COVID-MINDS repository](#).
- 8 For example, see Torous et al. (2018). Clinical review of user engagement with mental health smartphone apps: Evidence, theory and improvements. *Evidence Based Mental Health*, 21(3), 116–119. <https://doi.org/10.1136/eb-2018-102891>; and the consensus statement that followed: Torous et al. (2019). Towards a consensus around standards for smartphone apps and digital mental healthcare: Towards a consensus around standards for smartphone apps and digital mental healthcare. *World Psychiatry*, 18(1), 97–98. <https://doi.org/10.1002/wps.20592>

Endnotes

- 9 See the following notable publications: Phillips et al. (2021). Four Principles of Explainable Artificial Intelligence. National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.IR.8312>; Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1–38. <https://doi.org/10.1016/j.artint.2018.07.007>; Diakopoulos, N. (2015).
- 10 For a recent review of methods, see Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy*, 23(1), 18. <https://doi.org/10.3390/e23010018>
- 11 Ferreira, J. J., & Monteiro, M. S. (2020). What Are People Doing About XAI User Experience? A Survey on AI Explainability Research and Practice. In A. Marcus & E. Rosenzweig (Eds.), *Design, User Experience, and Usability. Design for Contemporary Interactive Environments* (Vol. 12201, pp. 56–73). Springer International Publishing. https://doi.org/10.1007/978-3-030-49760-6_4
- 12 Information Commissioner's Office & Alan Turing Institute. (2020). Explaining decisions made with AI. <https://ico.org.uk/media/for-organisations/guide-to-data-protection/key-data-protection-themes/explaining-decisions-made-with-artificial-intelligence-1-0.pdf>; Algorithmic Accountability: Journalistic investigation of computational power structures. *Digital Journalism*, 3(3), 398–415. <https://doi.org/10.1080/21670811.2014.976411>.
- 13 Information Commissioner's Office & Alan Turing Institute. (2020). Explaining decisions made with AI. <https://ico.org.uk/media/for-organisations/guide-to-data-protection/key-data-protection-themes/explaining-decisions-made-with-artificial-intelligence-1-0.pdf>
- 14 Ward, F. R., & Habli, I. (2020). An Assurance Case Pattern for the Interpretability of Machine Learning in Safety-Critical Systems. In A. Casimiro, F. Ortmeier, E. Schoitsch, F. Bitsch, & P. Ferreira (Eds.), *Computer Safety, Reliability, and Security. SAFECOMP 2020 Workshops* (Vol. 12235, pp. 395–407). Springer International Publishing. https://doi.org/10.1007/978-3-030-55583-2_30
- 15 Not to be confused with the related SAFE-D principle, 'Sustainability'.
- 16 ICO and Alan Turing Institute. Explaining decisions made with AI. Technical Report, Information Commissioners Office, May 2020. URL: <https://ico.org.uk/media/for-organisations/guide-to-data-protection/key-data-protection-themes/explaining-decisions-made-with-artificial-intelligence-1-0.pdf> (visited on 2020-10-26).
- 17 Christopher Burr and David Leslie. Ethical assurance: a practical approach to the responsible design, development, and deployment of data-driven technologies. *AI and Ethics*, June 2022. URL: <https://link.springer.com/10.1007/s43681-022-00178-0> (visited on 2022-08-01), [doi:10.1007/s43681-022-00178-0](https://doi.org/10.1007/s43681-022-00178-0).

Conclusion

- 1 Burr, C., and Leslie, D. (2022). Ethical assurance: a practical approach to the responsible design, development, and deployment of data-driven technologies. *AI Ethics*. <https://doi.org/10.1007/s43681-022-00178-0>
- 2 Details of the platform and the code is available through the following GitHub repository, and we welcome contributions to its ongoing development from any open-source developers: <https://github.com/alan-turing-institute/Assurance-Platform>



turing.ac.uk
@turinginst