




# Ethical assurance: a practical approach to the responsible design, development, and deployment of data-driven technologies

Christopher Burr<sup>1</sup> · David Leslie<sup>1</sup> 

Received: 20 October 2021 / Accepted: 21 May 2022 / Published online: 22 June 2022  
© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2022

## Abstract

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

**Keywords** Argument-based assurance · Ethical assurance · AI ethics · AI governance · Machine learning lifecycle

## 1 Introduction

The recent history of artificial intelligence (AI) ethics and governance has been characterised by increasingly vocal calls for a move from principles to practice. Over the past several years, some have discerned a rapid transition in the field from an initial concentration on high-level principles and techno-solutionist “fixes” (e.g. for issues such as algorithmic bias) towards a “third wave” of hard-nosed advocacy and legal action that is focused on “practical mechanisms for rectifying power imbalances and achieving individual and societal justice” [42]. Others have emphasised that “closing the gap” between principles and practice should involve the employment of myriad tools and methods throughout the

various stages of a project’s lifecycle, so that the “what” of ethical principles can be translated into the “how” of “technical mechanisms” [53]. Others still have called for a strengthening of regimes of “auditability,” “traceability,” and “reviewability,” emphasising the importance of oversight, accountability, and transparency as the key to the effective governance for responsible AI research and innovation [16, 43, 54, 60].

Notwithstanding the substantial merits of this intensifying concentration on the intersection of moral concepts and social praxis, these perspectives have fallen short of fully realising the transformations they identify and promote. Those who have turned to the incorporation of a patchwork of technical tools and documentation methods into the various stages of the AI or machine learning (AI/ML) project lifecycle have provided an important bird’s eye view of recording, auditing, and standards conformity desiderata. They have focused, for instance, on how to document the “creation, composition, intended uses, maintenance, and other properties” of datasets [27] or on how to “encourage transparent model reporting” through

✉ Christopher Burr  
cburr@turing.ac.uk

✉ David Leslie  
dleslie@turing.ac.uk

<sup>1</sup> The Alan Turing Institute, 96 Euston Road,  
London NW1 2DB, UK

“documentation detailing their performance characteristics” [51]. Such documentation-centred governance strategies, however, have run the risk of remaining too far above, and outside of, the actual sociotechnical processes behind the innovation practices they endeavour to document. The problem here is not that tools and method such as Datasheets [27], Data Nutrition Labels [36], Data Statements [7], Model Cards [51], and FactSheets [3] are of *no use* as provisional attempts at closing the gap between principles and practice, but rather that the limited perspective they take is liable to neglect the social, cultural, and cognitive preconditions of the responsible innovation practices they aspire to advance.

Cobbling together “a robust ‘toolbox’ of mechanisms to support the verification of claims about AI systems and development processes” [11], in this latter sense, leads, in AI ethics and governance, to a kind of functional tardiness of the governance strategies that result. Namely, it leads to an emphasis on narrowly targeted methods such as “effective assessment” [11], “auditability” [54, 60], “traceability” [43], and “reviewability” [16], that show up on the scene a moment too late. Such methods remain *ex post facto* and external to the inner workings of sufficiently reflective and responsible modes of technology production and use. It is at this latter, more foundational level of cultural formation, value shaping, and action orientation that a bridging of the gulf between principles and practice in AI ethics and governance must begin.

Beyond off-the-shelf tools and documentation-centred governance instruments, closing the gap between principles and practice requires a transformation of organisational cultures, technical approaches, and individual attitudes from inside the processes and practices of design, development, and deployment themselves. Achieving this requires researchers, technologists, and innovators to establish and maintain end-to-end habits of critical reflection and deliberation throughout all stages of a research or innovation project’s lifecycle. This more basic organisational, technical, and attitudinal transformation entails that designers and developers of data-driven technologies pay deliberate and continuous attention to the role that values play in both discovery and engineering processes as well as in considerations of the real-world effects that these processes yield. It requires sustained interdisciplinary efforts to consider the multi-faceted contexts of research and innovation, to anticipate potential impacts, and to engage affected stakeholders inclusively to ensure appropriate forms of social licence and democratic governance. In contrast, an approach to building trustworthy AI/ML systems that takes as its starting point a focus on technologically based tools or documentation protocols (like those mentioned above) works from the outside in, all while the actual change required to bridge the divide between principles and practice must instead originate from within actual research and innovation activities as part of a deeper transformation of the organisational environments

and individual attitudes, standpoints, and dispositions whence those activities derive. In this article, we propose a systematic and considered step towards this practice-driven and process-based approach to responsible AI/ML innovation by introducing a version of argument-based assurance that we call ‘ethical assurance.’

## 1.1 Article overview

For the purpose of this article, we offer the following definition of argument-based assurance (ABA):

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.

ABA is already widely used in safety-critical domains or industries where manufacturing and development processes are required to comply with strict regulatory standards and support industry-recognised best practices [34]. The output of this process is typically an assurance case, which can offer a formal, textual, or visual representation of the argument that seeks to demonstrate how regulatory goals or standards have been met. This process also supports related goals such as facilitating transparent communication and establishing trust between system or product developers and stakeholders.

In this paper, we seek to generalise the method of ABA to account for wider normative goals, related to ethical principles such as sustainability, accountability, fairness, or explainability. This generalised version, known as ‘ethical assurance’ provides a structured method for reflecting upon how and whether normative goals have been sufficiently established throughout the design, development, and deployment of an AI or data-driven technology, while also facilitating a process of active enquiry that supports meaningful stakeholder participation and deliberation. The participatory component is necessary for ensuring that the ethical assurance cases have moral legitimacy as well as social licence, and also helps to overcome concerns about so-called “ethics washing” [32] (see Sect. 5.1).

In generalising ABA to accommodate wider normative considerations, we offer a framework that can support anticipatory and reflective assessment of a project’s social commitments and responsibilities, outline a procedural method for operationalising ethical principles that result in justifiable forms of action-guidance, and address the practical needs of technical project governance for complex data-driven technologies.

The structure of our article is as follows. In Sect. 2, we introduce argument-based assurance in the context of AI/ML, and explain the current scope and limitations of existing research.

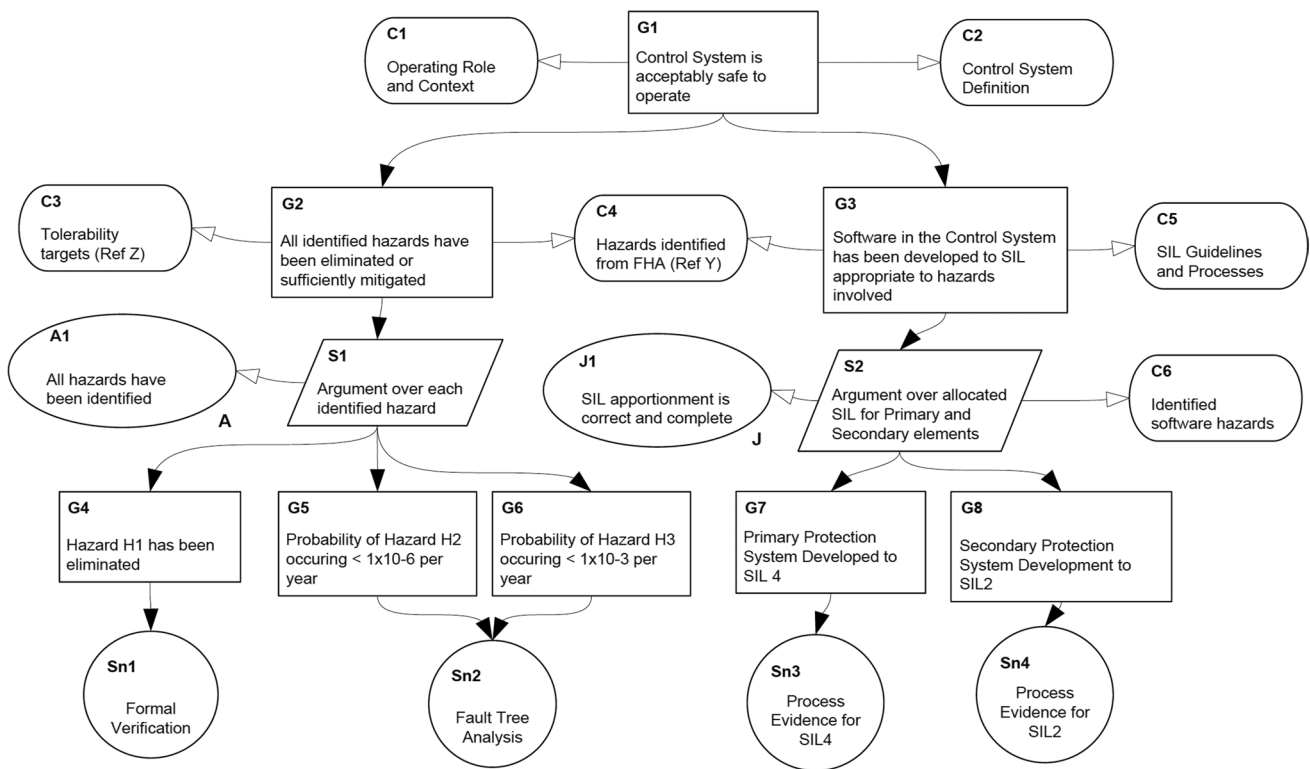


Fig. 1 An example assurance case focussed on safety of a control system (reprinted from [20])

In Sect. 3, we lay the foundations for our positive proposal, while also addressing some of the limitations of prior research. First, we present a heuristic model of the AI/ML project lifecycle that treats the design, development, and deployment of such technologies as sociotechnical constructs. Second, we show this model can serve as a form of scaffolding for a more reflective and participatory form of deliberation regarding assurance.

In Sect. 4, with the foundations set, we present the ethical assurance methodology.<sup>1</sup> We introduce the key elements required for building an ethical assurance case and discuss the procedural aspects that underpin the justifiability of normative claims made about sociotechnical systems. We then explore several topics related to ethical assurance, including how to evaluate evidence that is employed to justify normative, goal-directed claims, and why it is important to adopt a reflective and phased approach to the development of ethical assurance.

In Sect. 5, we conclude by anticipating and responding to several challenges that could be raised against ethical assurance, and identify several open issues and possible next steps for the project.

<sup>1</sup> Readers who wish to jump straight to our positive proposal can begin with this section, but in doing so will skip over important details that explain the context and motivation for the proposal itself.

## 2 Argument-based assurance

Assurance is a process of establishing trust. In safety-critical domains (e.g. automotive, energy), where trust is vital, and where manufacturing and development processes are both required to comply with strict regulatory standards and, ideally, reflect industry-recognised best practices, ABA is widespread. A common way to meet these compliance requirements is through the production of assurance cases, which provide a systematic method for justifying technical claims regarding specific properties of a system. An assurance case can be defined as follows:

“A reasoned and compelling argument, supported by a body of evidence, that a system, service or organisation will operate as intended for a defined application in a defined environment” ([20], 10).

Assurance cases tend to have a particular focus or goal. For example, the *safety case* in Fig. 1 is directed towards and structured around a clearly defined goal (G1, top of figure) of demonstrating that a control system is “acceptably safe to operate,” within a given operating role and context (C1) (e.g. a component in an aircraft that will be used in well-defined environments) [10]. Other assurance cases may focus on the security, availability, or maintenance of a system.

## 2.1 Assurance of machine learning and AI systems

Increasing concerns about the safe operation of autonomous, adaptive, and data-driven technology has resulted in a growing interest in the use of ABA for assessing and assuring ML or AI systems [4, 59, 70]. This research fits within a broader assurance ecosystem, which goes beyond the specific methodology of ABA to draw together myriad legal, ethical, and social concerns [14].

Within this broader remit, there is research that includes general overviews or frameworks that support transparent reporting and communication [11, 51], specific (narrowly focussed) tools that support bias mitigation or algorithmic interpretability [38, 50, 58, 63], as well as more focussed extensions of assurance cases to address the specific challenges of ML [4, 30, 70]. Each of these approaches can play a valuable role individually, but collectively add up to a (currently) disorganised toolbox of practical mechanisms with little unifying purpose or direction.

Brundage et al. [11] provide some means for bringing order to this miscellany by identifying practical mechanisms to support the trustworthy development of AI systems, which are categorised according to whether they are ‘institutional,’ ‘software,’ or ‘hardware’ mechanisms. For example, they note that institutional mechanisms, such as third party auditing can be used to create a “robust alternative to self-assessment claims,” while bias and safety bounties can “strengthen incentives to discover and report flaws in AI systems” (Brundage et al. [11], 1). Similarly, the software mechanisms they discuss also overlap with tools that support tasks such as bias mitigation or explainability. However, insofar as the purpose of this taxonomy is to help with the evaluation of “verifiable claims,” which the authors define as “statements for which evidence and arguments can be brought to bear on the likelihood of those claims being true,” the mechanisms themselves are insufficient.<sup>2</sup>

Turning to one of the most comprehensive proposals, Ashmore et al. [4] provide a systematic survey of ML assurance, focussing on the generation of evidential artefacts that can be used in the process of developing and evaluating an ML system. In a similar vein to Brundage et al. [11], their approach covers the methods and mechanisms that can provide evidence for claims about the properties of ML systems. However, their approach categorises these methods according to where in the ML lifecycle they are most relevant. As they note, each of the stages within this lifecycle have different desiderata that affect the generation

of evidential artefacts. For example, they argue that (from an assurance perspective) an ML model should exhibit the following properties: performant, robust, reusable, and interpretable.<sup>3</sup> In addition, for each of these key desiderata, there are various methods that can generate evidence to support corresponding claims (e.g. providing details of regularisation methods can verify claims about the robustness of the model, and transfer learning can support reusability claims).

The survey of methods that Ashmore et al. [4] present is impressive, going beyond technical goals such as safety or reliability. This work has also, more recently, been extended by several of the original authors to connect it more directly to ABA (see [34]) and also demonstrate how it could support the assurance of AI explainability in domains such as health-care. For example, Ward and Habli [70] develop a model template, known as an ‘argument pattern’, for assuring the interpretability of ML systems.

Interpretability is a key desiderata discussed in Ashmore et al. [4] and a vital component in recent efforts to improve the explainability of AI systems [38]. Therefore, as Ward and Habli [70] argue, providing assurance that a particular model is “sufficiently interpretable” in a given context (i.e. a particular time, setting, and a specified audience), helps build confidence in the use of the system, and thus has ethical significance. However, while the pattern that Ward and Habli develop is generalisable to a range of contexts, it is nevertheless framed in terms of safety concerns.<sup>4</sup> That is, the reason for assuring a system’s interpretability is grounded in the necessity of demonstrating that it is ‘safe to operate.’ While this has the effect of anchoring requirements, such as interpretability, in clearly articulated safety outcomes, it simultaneously divorces it from wider normative considerations that are captured by more inclusive goals such as explainability [39], or related principles, such as respect for autonomy or informed consent.<sup>5</sup>

<sup>2</sup> It is important to acknowledge that [11] recognise that the mechanisms alone are merely tools to support wider processes of governance, and also suggest the need for pursuing argument-based forms of assurance in Appendix III.

<sup>3</sup> Ashmore et al. [4] also define key desiderata for each of the four stages of their “ML lifecycle”: data management, model learning, model verification, and model deployment.

<sup>4</sup> For instance, consider the following statement from (Hawkins et al. [34], 13): “requirements such as security or usability should be defined as ML safety requirements only if the behaviours or constraints captured by these requirements influence the safety criticality of the ML output. ‘Soft constraints’ such as interpretability may be crucial to the acceptance of an ML component especially where the system is part of a socio-technical solution. All such constraints defined as ML safety requirements must be clearly linked to safety outcomes.”

<sup>5</sup> Ward and Habli do acknowledge that the first step in the process of developing an assurance case centred upon interpretability is to “ask why the project needs interpretability and set the desired requirements that the project should satisfy.” Therefore, it is possible that the pattern they offer may also serve to provide assurance for wider (interpretability-linked) normative goals.

These previous examples offer valuable and worthwhile contributions to the current literature on AI assurance. However, the existing literature is nevertheless limited in several ways, all of which are related to its *scope*.

First, the proposed frameworks, which claim to be “end-to-end,” often do not make sufficient room for wider normative considerations that arise prior to stages such as data extraction or model development. For instance, a growing literature has drawn attention to the existence of historical or social biases that affect the fairness, validity, performance, and trustworthiness of ML systems [8, 9, 40, 65]. The neglect of such considerations (e.g. the existence of historical patterns of social discrimination) can often cause cascading effects through the ML lifecycle, if not properly considered or addressed [46].

Second, the current literature is often too narrowly focussed on technological solutions (e.g. FairML “solutions” to complex social justice issues). However, many social problems require a broader, more nuanced, and deliberative approach, rarely reducible to a single solution. Rather, as seen in the recent failure of many contact-tracing systems, it is the absence of a legitimating social licence that often leads to finite public resources going to waste [45].

Finally, and perhaps most importantly for the present paper, the current literature is focussed on a limited set of goals, such as safety or security. While this is understandable in a technical industry focused on regulatory compliance, the myopic focus on merely doing what is necessary, rather than what is best, is often unsustainable in the long-term. The market dynamics that may emerge from such a collective attitude can generate a “race to the bottom,” as has been seen recently in the domain of digital marketing and advertising (e.g. data privacy scandals).

These above limitations need to be addressed if we are to fully embed ethical considerations into the design, development, and deployment of ML. Therefore, in laying out our positive proposal we overcome these limitations, respectively, by (a) presenting a model of the AI/ML project lifecycle that is designed to support a more reflective and anticipatory form of assurance that addresses wider normative goals, (b) showing how this sociotechnical perspective can support the operationalisation of normative principles through a more inclusive and participatory form of deliberation, and (c) developing a generalisable method of ethical assurance that can extend the scope of current research.

### 3 Laying the foundations

In this section, we lay the foundations for our methodology, while addressing the first two limitations highlighted at the end of the last section.

#### 3.1 A sociotechnical approach to the AI/ML project lifecycle

There are many ways of carving up a project lifecycle for some data-driven technology (hereafter shortened to just ‘project lifecycle’). For instance, Sweenor et al. [68], who are focussed on machine learning operations (MLOps),<sup>6</sup> break it into four stages: build, manage, deploy and integrate, and monitor. Similarly, Ashmore et al. [4] identify four stages, which have a more specific focus on data science: data management, model learning, model verification, and model deployment. Furthermore, there are also well-established methods that seek to govern common tasks within a project lifecycle, such as data mining (e.g. CRISP-DM or SEMMA).

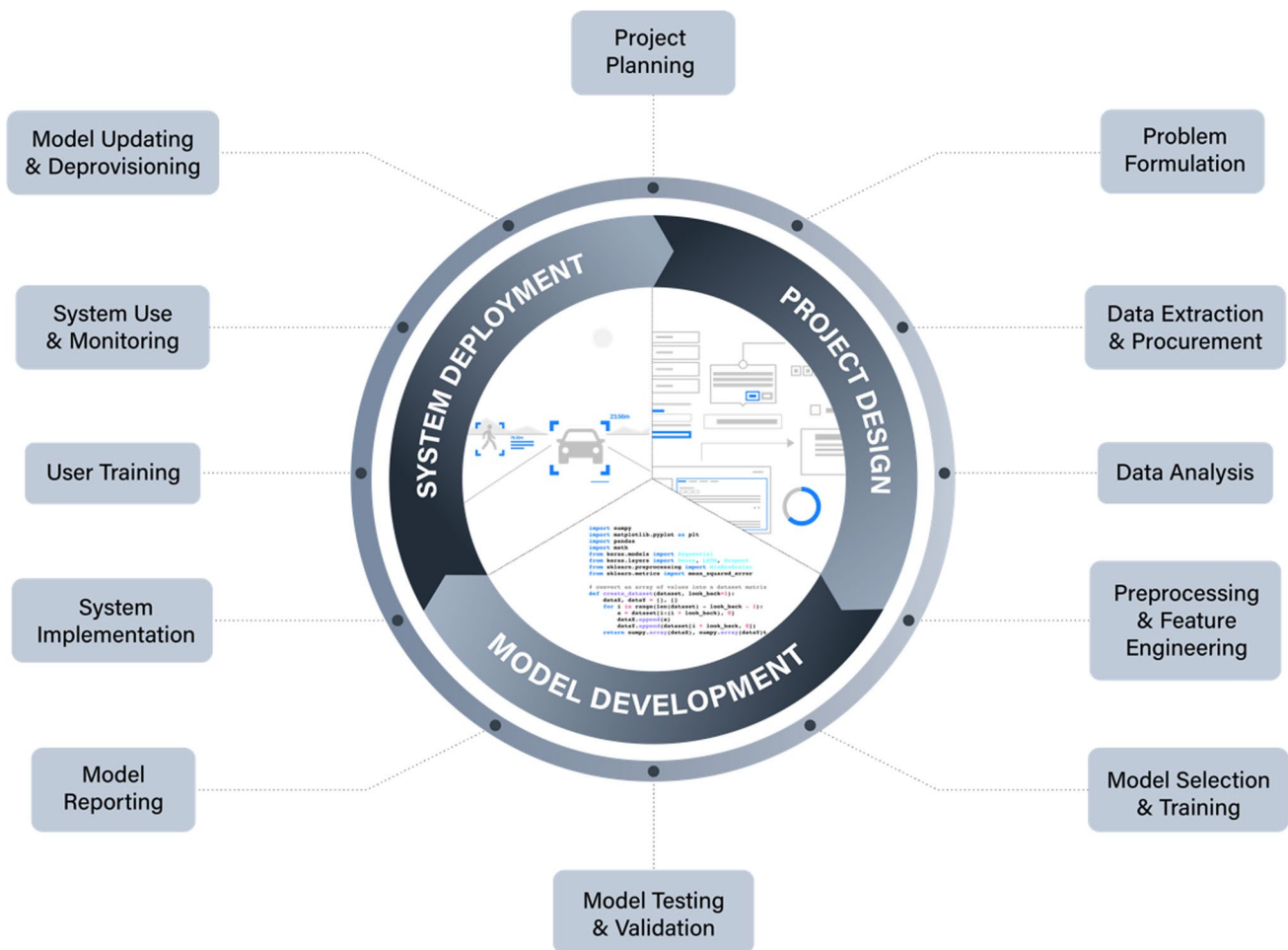
The multiplicity of approaches is likely a product of the evolution of diverse methods in data mining/analytics, the significant impact of ML on research and innovation, and the specific practices and considerations inherent to each of the various domains where ML techniques are applied. Since we are interested in developing a methodology of ABA that can address wider ethical and social concerns, therefore, it is important to have a sufficient understanding of what activities ought to fall within the scope of the project lifecycle, and, therefore, under the responsibility of the project team.

Figure 2 presents a model of the ML lifecycle that has been designed specifically to support the assurance process. Like other models, it (a) remains faithful to the importance of technical requirements and challenges, but also (b) specifically supports a more open, reflective, and participatory form of deliberation. In short, it can be thought of as a heuristic device to help scaffold and structure responsible forms of project governance.

To begin, the inner circle breaks the project lifecycle into a process of (project) design, (model) development, and (system) deployment. These terms are intended to be maximally inclusive. For instance, the design stage encompasses any project task or decision-making process that sets downstream constraints (e.g. design system constraints). Importantly, this includes ethical, social, and legal constraints, which we will discuss later.

The boundaries between these more general stages are not strict. Rather, each of the stages shades into its neighbours because in practice there is often no clearly delineated boundary that differentiates certain project design activities

<sup>6</sup> The term ‘MLOps’ refers to the application of DevOps practices to ML pipelines. The term is often used in an inclusive manner to incorporate traditional statistical or data science practices that support the ML lifecycle, but are not themselves constitutive of machine learning (e.g. exploratory data analysis), as well as deployment practices that are important within business and operational contexts (e.g. monitoring KPIs).



**Fig. 2** The project lifecycle—the overarching stages of design, development, and deployment for a typical data-driven project (e.g. ML algorithm or AI system)

(e.g. data extraction and exploratory analysis) from model design activities (e.g. data preprocessing, feature engineering, and model selection). As such, the design stage overlaps with the development stage, but the latter extends to include the actual process of training, testing, and validating a ML model. Similarly, the process of putting a model into production and implementing the encompassing system into an operable environment (i.e. system implementation) can be thought of as both a development and deployment activity. In addition, so, the deployment stage overlaps with the ‘development’ stage and the ‘design’ stage—the deployment of a system should be thought of as an ongoing process (e.g. where new data are used to continuously train the ML model, or, the decision to de-provision a model may require the planning and design of a new model, if the older (legacy) system becomes outdated).

Each higher level stage subsumes a wide variety of tasks and activities that are likely to be carried out by different individuals, teams, and organisations, depending on their

specific roles and responsibilities (e.g. procurement of data). Therefore, it is important to break each of the three higher level stages into their (typical) constituent parts, which are likely to vary to some extent between specific projects or within particular organisations. In doing so, we expose a wide range of diverse tasks, which give rise to a variety of ethical, social, and legal challenges.

The following sections provide an illustrative overview of these stages to help demonstrate how our model can go beyond the limitations of previous research and encapsulate wider normative considerations and tasks—the following is a non-exhaustive sample of the associated challenges. A summary of the stages is also presented in Table 1 as a reference.

### 3.1.1 (Project) design tasks and processes

*Project planning* Rather than using AI/ML as a “hammer” to go looking for nails, it is best to have a clear idea in mind

**Table 1** A summary of the project lifecycle stages

Project lifecycle activity	Summary 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 overarching 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 and evaluation metrics)
System implementation	The process of putting a model into production, and implementing the operational system, which enables and structures interaction with the model, within the respective environment (e.g. a recommender system that converts a user's existing movie ratings into recommendations for future watches)
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. consumers)
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 or deprovisioning (i.e. removing from the production environment)

of what the project's goals are at the outset. This can help to avoid a myopic focus on a narrow class of AI/ML-based "solutions," and also helps create space for a diversity of approaches—some of which may not require AI/ML at all. Project planning, therefore, can comprise a wide variety of tasks, including, but not limited to:

- an assessment of whether building an algorithmic model is the right approach given available resources and data, existing technologies and processes already in place, the complexity of the use-contexts involved, and the nature of the policy or social problem that needs to be solved [48];
- an analysis of user needs in relation to the prospective AI model and whether a solution involving the latter provides appropriate affordances in keeping with user needs and related functional desiderata;
- a contextual assessment of the target domain and of the expectations, norms, and requirements that derive therefrom;
- stakeholder analysis and team positionality reflection to determine the appropriate level and scope of community engagement activities [47];
- stakeholder impact assessment, supported by affected people and communities, to identify and evaluate pos-

sible harms and benefits associated with the project (e.g. socioeconomic inequalities that may be exacerbated as a result of carrying out the project), to gain social licence and public trust, and also feed into the process of problem formulation in the next stage;

- wider impact assessments—both where required by statute and done voluntarily for transparency and best practice (e.g. equality impact assessments, data protection impact assessments, human rights impact assessment, and bias assessment)

*Problem formulation* Here, 'problem' refers both to a well-defined computational process (or a higher level abstraction of the process) that is carried out by the algorithm to map inputs to outputs and to the wider practical, social, or policy issue that will be addressed through the translation of that issue into the statistical or mathematical frame. For instance, on the computational side, a convolutional neural network carries out a series of successive transformations by taking (as input) an image, encoded as an array, to produce (as output) a decision about whether some object is present in the image. On the practical, social, and policy side, there will be a need to define the computational "problem" being solved in terms of the algorithmic system's embeddedness in the social environment and to explain how

it contributes to (or affects) the wider sociotechnical issue being considered. In the convolutional neural network example, the system being produced may be a facial recognition technology that responds to a perceived need for the biometric identification of criminal suspects by matching face images in a police database. The social issue of wanting to identify suspects is, in this case, translated into the computational mechanism of the computer vision system. But, beyond this, diligent consideration of the practical, social, or policy issue being addressed by the system will also trigger, *inter alia*, reflection on the complex intersection of potential algorithmic bias, the cascading effects of sociohistorical patterns of racism and discrimination, wider societal and community impacts, and the potential effects of the use of the model on the actors in the criminal justice systems who will become implementers and subjects of the technology.

Sociotechnical considerations are also important for determining and evaluating the choice of target variables used by the algorithm, which may ultimately be implemented within a larger automated decision-making system (e.g. in a verification system). The task of formulating the problem allows the project team to get clear on what input data will be needed, for what purpose, and whether there exists any representational issues in, for example, how the target variables are defined. It also allows for a project team (and impacted stakeholders) to reflect on the reasonableness of the measurable proxy that is used as a mathematical expression of the target variable, for instance, whether being taken into care within 6 months of a visit from child protective services is a reasonable proxy for a child's being "at risk" in a predictive risk model for children's social care. The semantic openness and contestability of formulating problems and defining target variables in AI/ML innovation lifecycles is why stakeholder engagement, which helps bring a diversity of perspectives to project design, is so vital, and why this stage is so closely connected with the interpretive burdens of the project planning stage (e.g. discussion about legal and ethical concerns regarding permissible uses of personal or sensitive information).

*Data extraction or procurement* Ideally, the project team should have a clear idea in mind (from the planning and problem formulation stages) of what data are needed prior to extracting or procuring them. This can help mitigate risks associated with over-collection of data (e.g. increased privacy or security concerns) and help align the project with values such as data minimisation [39]. Of course, this stage may need to be revisited after carrying out subsequent tasks (e.g. preprocessing, model testing) if it is clear that insufficient or imbalanced data were collected to achieve the project's goals. Where data are procured, questions about provenance arise (e.g. legal issues, concerns about informed consent of human data subjects). Generally, responsible data extraction and procurement require the incorporation

of domain expertise into decision-making so that desiderata of data minimisation as well as of securing relevant and sufficient data can be integrated into design choices.

*Data analysis* Exploratory data analysis is an important stage for hypothesis generation or uncovering possible limitations of the dataset that can arise from missing data, in turn identifying the need for any subsequent augmentation of the dataset to deal with possible class imbalances. However, there are also risks that stem from cognitive biases (e.g. confirmation bias) that can create cascading effects that effect downstream tasks (e.g. model reporting).

### 3.1.2 (Model) development tasks and processes

*Preprocessing and feature engineering* Preprocessing and feature engineering is a vital but often lengthy process, which overlaps with the design tasks in the previous section and shares with them the potential for human choices to introduce biases and discriminatory patterns into the AI/ML workflow. Tasks at this stage include data cleaning, data wrangling or normalisation, and data reduction or augmentation. It is well understood that the methods employed for each of these tasks can have a significant impact on the model's performance (e.g. deletion of rows versus imputation methods for handling missing data). As Ashmore et al. [4] note, there are also various desiderata that motivate the tasks, such as the need to ensure the dataset that will feed into the subsequent stages is relevant, complete, balanced, and accurate. At this stage, human decisions about how to group or disaggregate input features (e.g. how to carve up categories of gender or ethnic groups), or about which input features to exclude altogether (e.g. leaving out deprivation indicators in a predictive model for clinical diagnostics), can have significant downstream influences on the fairness and equity of an AI/ML system.

*Model selection* This stage determines the model type and structure that will be produced in the next stages. In some projects, model selection will result in multiple models for the purpose of comparison based on some performance metric (e.g. accuracy). In other projects, there may be a need to first implement a pre-existing set of formal models into code. The class of relevant models is likely to have been highly constrained by many of the previous stages (e.g. available resources and skills, problem formulation), for instance, where the problem demands a supervised learning algorithm instead of an unsupervised learning algorithm; or where explainability considerations require a more interpretable model (e.g. a decision tree).

*Model training* Prior to training the model, the dataset will need to be split into training and testing sets to avoid model overfitting. The training set is used to fit the ML model, whereas the testing set is a hold-out sample that is used to evaluate the fit of the ML model to the underlying



data distribution. There are various methods for splitting a dataset into these components, which are widely available in popular package libraries (e.g. the scikit-learn library for the Python programming language). Again, human decision-making at this stage about the training–testing split and about how this shapes desiderata for external validation—a subsequent process where the model is validated in wholly new environments—can be very consequential for the trustworthiness and reasonableness of the development phase of an AI/ML system.

*Model validation and testing* The testing set is typically kept separate from the training set, to provide an unbiased evaluation of the final model fit on the training dataset. However, the training set can be further split to create a validation set, which can then be used to evaluate the model while also tuning model hyperparameters. This process can be performed repeatedly, in a technique known as (k-fold) cross-validation, where the training data are resampled (*k*-times) to compare models and estimate their performance in general when used to make predictions on unseen data. This type of validation is also known as ‘internal validation,’ to distinguish it from external validation, and, in a similar way to choices made about the training–testing split, the way it is approached can have critical consequences for how the performance of a system is measured against the real-world conditions that it will face when operating “in the wild.”

*Model reporting* Although the previous stages are likely to create a series of artefacts while undertaking the tasks themselves, model reporting should also be handled as a separate stage to ensure that the project team reflect on the future needs of various stakeholders and end users. While this stage is likely to include information about the performance measures used for evaluating the model (e.g. decision thresholds for classifiers, accuracy metrics), it can (and should) include wider considerations, such as intended use of the model, details of the features used, training–testing distributions, and any ethical considerations that arise from these decisions (e.g. fairness constraints, use of politically sensitive demographic features).<sup>7</sup>

### 3.1.3 (System) Deployment tasks and processes

*System implementation* Unless the end result of the project is the model itself, which is perhaps more common in scientific research, it is likely that the model will need to be embedded within a larger system. This process, sometimes known as ‘model productionalisation,’ requires understanding (a) how

the model is intended to function in the proximate system (e.g. within an agricultural decision support system used to predict crop yield and quality) and (b) how the model will impact—and be impacted by—the functioning of the wider sociotechnical environment that the tool is embedded within (e.g. a decision support tool used in healthcare for patient triaging that may exacerbate existing health inequalities within the wider community). Ensuring the model works within the proximate system can be a complex programming and software engineering task, especially if it is expected that the model will be updated continuously in its runtime environment. But, more importantly, understanding how to ensure the system’s sustainability given its embeddedness in complex and changing sociotechnical environments requires active and contextually informed monitoring, situational awareness, and vigilant responsiveness.

*User training* Although the performance of the model is evaluated in earlier stages, the model’s impact cannot be entirely evaluated without consideration of the human factors that affect its performance in real-world settings. The impact of human cognitive biases, such as algorithmic aversion<sup>8</sup> must also be considered, as such biases can lead to over- and under-reliance on the model (or system), in turn negating any potential benefits that may arise from its use. Understanding the social and environmental context is also vital, as sociocultural norms may contribute to how training is received, and how the system itself is evaluated (see [12]).

*System use and monitoring* Depending on the context of deployment, it is likely that the performance of the model could degrade over time. This process of model drift is typically caused by increasing variation between how representative the training dataset was at the time of development and how representative it is at later stages, perhaps due to changing social norms (e.g. changing patterns of consumer spending, evolving linguistic norms that affect word embeddings). As such, mechanisms for monitoring the model’s performance should be instantiated within the system’s runtime protocols to track model drift, and key thresholds should be determined at early stages of a project (e.g. during project planning or in initial impact assessment) and revised as necessary based on monitoring of the system’s use.

*Model updating or deprovisioning* As noted previously, model updating can occur continuously if the architecture of the system and context of its use allows for it. Otherwise, updating the model may require either revisiting previous stages to make planned adjustments (e.g. model selection and training), or if more significant alterations are required

<sup>7</sup> There is some notable overlap between this stage of the project lifecycle and the ethical assurance methodology, as some approaches to model reporting often contain similar information that is used in building an ethical assurance case [4, 51], specifically in the process of establishing evidential claims and warrant (see Sect. 4.2).

<sup>8</sup> Algorithmic aversion refers to the reluctance of human agents to incorporate algorithmic tools as part of their decision-making processes due to misaligned expectations of the algorithm’s performance (see [12]).

the extant model may need to be entirely de-provisioned, necessitating a return to a new round of project planning.

This overview and summary of the project lifecycle, summarised in Table 1, is by necessity an abstraction. However, it provides a useful anchor for subsequent discussion, and serves to motivate the following question: how do you provide assurance for the diversity of tasks included throughout the process? For instance, there may be a plurality of ethical goals relevant to the assurance of (model) development or system use and monitoring, including demonstrating that the system being deployed is safe, secure, fair, trustworthy, explainable, sustainable, or respectful of human agency and autonomy. How do you provide assurance that the interconnected project processes and activities individually and collectively support the relevant goal? This is why we need a unifying framework and methodology that makes space for the operationalisation of end-to-end, normative considerations and complements existing regulatory culture, as opposed to merely a miscellany of practical mechanisms.

With this model of a more sociotechnical approach to design, development, and deployment outlined, we now turn to explain how teams can identify, specify and operationalise ethical values and principles.

### 3.2 The normative foundations of ethical assurance

There is much confusion about the role that normative concepts play in supporting ethical reasoning and decision-making. A common set of worries hinge upon the misconception that ethical values and principles, such as respect for autonomy or fairness, are too vague or abstract to be actionable in a technical context (see Sect. 1), can disguise deep-seated disagreements, and even enable ethics washing. These worries are borne, however, from a misunderstanding of the role that ethical values and principles are designed to play in ethical deliberation. In brief, while values and principles are not sufficient in themselves for action-guidance, they do play a vital, contributory, and sometimes explanatory or justificatory role in deliberation. In this section, we argue that they are best viewed as starting points in an ongoing, participatory process of reflection, action, and justification.<sup>9</sup>

To begin with, ethical values can support a practical process of anticipatory reflection during the planning stages of a project by offering normative criteria against which the potential adverse impacts of prospective AI/ML systems can be assessed. In this way, they can provide common starting points for scoping, identifying and evaluating the ways in which individual and communities could be positively or

negatively affected by the deployment of a prospective system. The reason why values should operate here as departure points for reflection rather than as fixed and predetermined directives for action is that they are contextually bounded, open to continuous re-interpretation and, therefore, always subject to ongoing discursive negotiation. Values and principles may mean very different things to different people, and demand very different actions in the varying contexts of different projects.

For example, consider how the ethical value of ‘respecting individual autonomy’ will mean very different things for two projects that aim to develop an AI chatbot, where one supports the education of children and the other is used to support the assessment of individuals suffering from mental health disorders. That is, the value is specified in different ways for the two projects, despite serving as a fundamental standard of conduct from which other related moral standards or judgement are drawn. For instance, a healthcare professional may recognise that restricting an individual’s freedom, if they are suffering from a severe mental health disorder, is necessary, while nevertheless doing so in a way that respects their individual autonomy during care and treatment. In this way, ethical values and principles may be binding, but only as *pro tanto* reasons for action—that is, reasons that speak in favour of some goal or claim but that can be overridden by additional, competing reasons. Because of this, it is important to use a process of critical and inclusive reflection not only to identify the relevant values and principles that underpin a project and that provide a normative compass for evaluation of its potential impacts, but also to identify the different interpretations that affected stakeholders may have regarding the values and principles in question.

This point speaks to the (admittedly misplaced) confusion that is directed towards the role of principles in the domain of technology governance. General principles do not fully determine actions or judgements; their substantive content is insufficient for directing action without first addressing how the principle is specified in a particular context (e.g. the principle of ‘transparency’ may have a narrower meaning in the context of criminal justice or healthcare, where disclosure of sensitive information ought to be restricted, than it does in the context of manufacturing or agriculture, where no personal data are used). This sort of contextual and interpretive responsiveness requires additional reflection and deliberation, often with domain experts and stakeholders. Hence, principles should be treated, as Beauchamp and Childress [5] note, “less as firm directives that are applied and more as general guidelines that are explicated and made suitable for specific tasks, as often occurs in formulating policies and altering practices.”

Because of this, principles can be supported and complemented by additional processes that enable reflective

<sup>9</sup> This view derives, primarily, from the ethical theory of principlism [6], but is also reflected in contemporary research in responsible research and innovation [57].

and discursive practices (e.g. the processes of ethical assurance that we set out in the next section). Well-established self-evaluative mechanisms, for example, can help ensure that any actions taken throughout a project are in alignment with the general guidance of relevant principles (e.g. end-to-end bias self-assessment processes can help to animate and operationalise broadly accepted principles of fairness and non-discrimination). In addition, practical mechanisms such as well-designed stakeholder impact assessments can ensure that possible harms to different groups are identified early on. In addition, the integration of regulatory guidance or legal precedents into contextual consideration about the reasonable expectations of affected people can constrain decision-making in positive ways, rather than merely restricting or inhibiting innovation.

This understanding of the complementary and socially embedded role of ethical values and principles as supporting and guiding practical decision-making is sometimes known as the “reflective equilibrium” model [61]—referring to the stable point of a deliberative process in which a group of individuals (or society more broadly) reflects on and revises a moral belief or judgement. The resulting judgement aims at maximising coherence among the linked set of beliefs that underpin the judgement, such that the holding of the judgement, while defeasible, is nevertheless justified (or, warranted) given the deliberative process. Of course, to be legitimate in the first place, this process of deliberation has to adhere to additional standards that are determined by the equity of the communicative context and that are demanded as normative preconditions of discursive will formation (e.g. following a rational procedure; ensuring participatory parity; redressing power imbalances that can distort intersubjective communication; representing a fair democratic process) [5, 28].

The previous discussion regarding the role of ethical values and principles belies an enormous complexity, which is well beyond the scope of this paper to explore. In closing, we wish to highlight a few salient details that will bridge this discussion of normative concepts with our constructive proposal for ethical assurance. Even if these points help to address some of the confusion about the role that ethical values and principles are intended to play, some practitioners may remain unsatisfied—especially with regard to anticipatory component of collaborative reflection, where critics remain sceptical about the possibility of effective foresight (i.e. the charge that none of us, in the end, possesses a “crystal ball”). For instance, as Raji et al. ([60], 33) note, “[...] it remains challenging for practitioners to identify the harmful repercussions of their own systems prior to deployment, and, once deployed, emergent issues can become difficult or impossible to trace back to their source.” This challenge is sometimes referred to as Collingridge’s dilemma, named

after David Collingridge [17], and is well known in the area of responsible research and innovation [57].

Ultimately, a process of ongoing *ex ante* reflection across the AI/ML innovation lifecycle is not intended to be a vessel of unassailable prediction. Rather, it is meant to institute structured practices of preventive deliberation and preemptive action that steward the identification and mitigation of the risk of harms associated with systems prior to their deployment, for instance, ensuring that potential harms to marginalised groups are identified through an inclusive and participatory process of stakeholder engagement to safeguard that easily identifiable issues do not go unaddressed. Ethical values and principles function, in this sense, as normative headlights that provide innovators and members of affected communities with the moral vocabularies and conceptual resources that are needed both to envisage possible moral worlds and to cooperatively articulate what good and bad looks like in those worlds so that they can be shaped in accordance with shared visions of a better tomorrow.

In spite of this, a degree of epistemic humility is invariably also required, as some downstream hazards may be impossible to envision in advance and thus to forestall entirely. For example, harms could result from difficult-to-detect cumulative or aggregate impacts that are often imperceptible and incremental, or they can emerge from unintended consequences—especially with novel technologies that shape our complex sociotechnical environments in unpredictable ways.<sup>10</sup> What is important here is that practical measures are taken to promote best practices that get out ahead of hazards arising from modifiable risk factors and that accordingly reduce the universe of possible harms. This means building in space for anticipatory reflection throughout the processes of design, development, and deployment, and building in time to return to previous steps if issues are only identified at a later stage in the project lifecycle.

In the next section, we will discuss how the process of ethical assurance accommodates these requirements throughout the lifecycle of a project.

## 4 Ethical assurance

We now turn to introduce our positive proposal: ethical assurance. This section is intended to provide an overview of the methodology but should not be treated as a ‘user guide’ for building an ethical assurance case.

<sup>10</sup> This does not mean that if a system has been deployed with little to no oversight, nor with any due consideration given to the transparency and accountability of the processes, and ends up causing significant harm, that those responsible should be able to claim that it was due to “unforeseeable risk.”

## 4.1 What is ethical assurance?

“an oral or written expression is a standpoint if it expresses a certain positive or negative position with respect to a proposition, thereby making it plain what the speaker or writer stands for. In addition, a series of utterances constitutes an argumentation only if these expressions are jointly used in an attempt to justify or refute a proposition, meaning that they can be seen as a concerted effort to defend a standpoint in such a way that the other party is convinced of its acceptability.” (Eemeren and Grootendorst [24], 3)

The above quotation from van Eemeren and Grootendorst [24] sets an important anchoring point for the following section, by making explicit that an ethical assurance case is an argument that seeks to justify a particular standpoint towards some proposition or claim (i.e. the ethical goal).

As such, ethical assurance is a type of ABA that retains much of the original approach but extends it to incorporate wider ethical concerns into the design, development, and deployment of data-driven technologies, such as AI/ML, in a systematic and pragmatic manner. This is reflected in the following definition, which retains much of our earlier definition of ABA:

Ethical assurance is a process of using structured argumentation to provide reviewable assurance that a particular set of normative claims about the corresponding ethical properties of a system are warranted given the available evidence.

Despite the similarities, an extension of ABA is needed because the traditional focus of assurance cases (e.g. on safety and reliability), while important and vital, is necessarily limited due to the alternative goals (e.g. compliance with technical standards). For instance, consider how a power station can operate safely and reliably in certain respects, while nevertheless harming the environment through pollution [60].<sup>11</sup> Furthermore, as the explicit inclusion of reviewability (or, contestability) suggests, ethical assurance supports and promotes a culture of active enquiry, necessary for ensuring the moral legitimacy of a project [55].

<sup>11</sup> Harm to the environment can of course be incorporated into a broader notion of ‘safety,’ such that pollution generated in the everyday operations of a power station are factored into a safety assessment. However, the point we wish to address here is that the scope of concepts, such as ‘safety’ and ‘reliability’ tends to reflect a domain-specific focus or set of priorities (e.g. compliance with technical or legal standards, rather than ethical principles).

## 4.2 The structure and elements of an ethical assurance case

The general structure of ethical assurance, or the building of an ethical assurance case more specifically, can be described as an iterative and cyclical process of reflection, action, and justification throughout the stages of the project lifecycle that are outlined in Sect. 3.1.

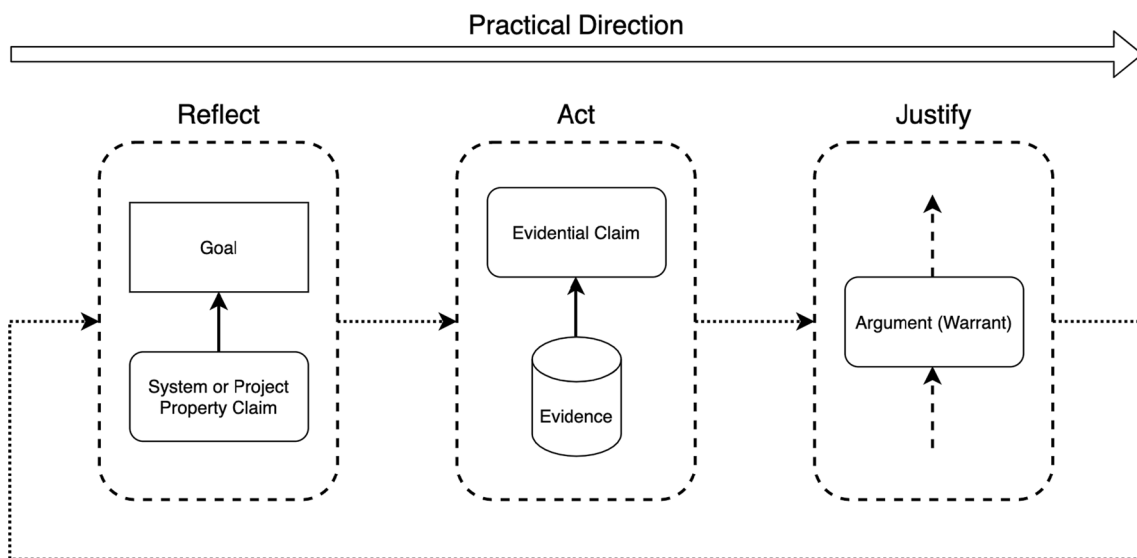
This iterative process involves (a) establishing the normative goals that identify and articulate key ethical qualities and determining the properties needed to assure these goals, (b) taking actions to operationalise these properties, and (c) compiling the evidence of these actions that then provides warrant for claims that the goals have been ascertained. The process of developing an assurance case assists internal reflection and deliberation, promoting the adoption of best practices and integrating these into design, development, and deployment lifecycles.

This is captured in Fig. 3, which presents a high-level schematic of the overarching process of reflect, act, justify, while also capturing the various elements that are required in an ethical assurance case:

- Top-level normative goal
- System or project property claims
- Argument (warrant)
- Evidential claim
- Evidence

This schematic follows the practical direction of project design, model development, and system deployment. For instance, a project team may begin in the project planning stage by identifying the normative goal of sustainability, which requires *inter alia* that the practices behind the system’s production and use be informed by ongoing considerations of the potential for exposing affected people and groups to harmful impacts. At the project design phase, operationalising this goal will involve engaging in anticipatory deliberations about the potential impacts of the project on the individuals and communities it could affect. Such a process of impact assessment plays a vital role in helping to set the direction of travel for the project (e.g. ensuring that their system protects fundamental human rights and freedoms and prioritises social justice) and in providing a shared vocabulary for project team appraisal, stakeholder engagement, and public communication. The action taken by the project team to realise the goal of sustainability at the project planning stage, namely the initial impact assessment provides evidence to justify the claim that in the project design phase, the goal of sustainability has been ascertained.

Sustainability is, of course, just one of many top-level normative goals that may be identified and articulated by project teams as they reflect on the salient ethical principles



**Fig. 3** A high-level schematic depicting the process of building an ethical assurance case. The stages of reflect, act, and justify are connected to the elements of an ethical assurance case

needed to assure the responsibility of their AI system or ML model and the trustworthiness of the practices behind building and using it. Key values associated with AI ethics also help with the specification of the ethical principles that serve as the top-level normative goals for the project—akin to the goal of safety in a traditional assurance case. With these goals set, the project team can then identify the necessary properties of the project or system that must be established throughout the design, development, and deployment of the system, and subsequently assured to justify how the relevant goal has been obtained.

This anticipatory process supports a responsible and ethical approach to project governance, but also speeds up the process of building an ethical assurance case. The decisions and actions that the project team take at each stage of the project lifecycle will invariably alter and refine their initial, anticipatory reflections, requiring a cyclical and interactive process of reflection, action, and justification. However, having a rough idea of what is required to ensure, say, that the fairness or the explainability of a system is sufficiently established can support a phased approach to building an assurance case, instead of leaving it to the final part of the project.

The following sections will follow this practical direction of building an assurance case. To illustrate the role of each element within an ethical assurance case, we will use the running example of a hypothetical project to design and develop a decision support tool to be deployed in a healthcare setting.

---

Illustrative case study

---

A decision support tool that uses a ML algorithm to triage (classify) incoming patients on the basis of their observable symptoms and physiological measurements ( $X$ ), to determine their expected risk of clinical deterioration ( $Y$ ),<sup>12</sup> and offer tailored guidance to the relevant healthcare practitioner<sup>13</sup>

---

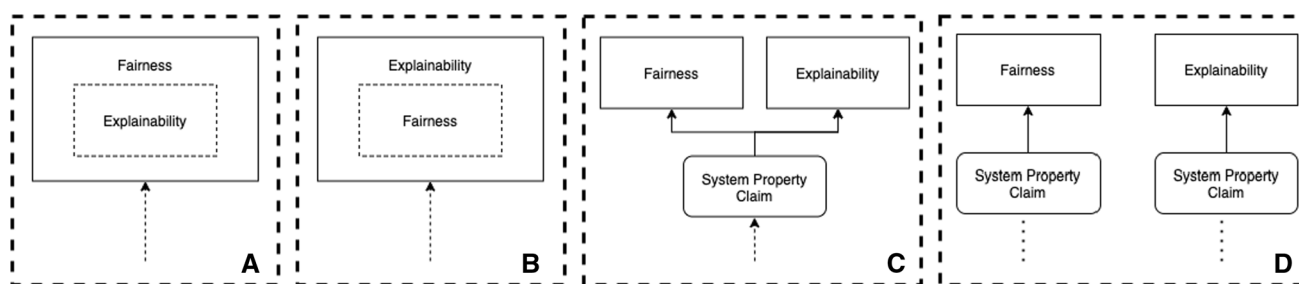
**4.2.1 Top-level normative goal**

An ethical assurance case begins with the identification of top-level normative goals, which are oriented towards key ethical values (see Sect. 3.2).

For example, our hypothetical project team are motivated (in part) by the recognition that their tool could lead to an unequal and unfair distribution of health outcomes among their target population, in virtue of (a) its discriminatory ability to classify patients (e.g. based on demographic,

<sup>12</sup> In formal terms, we can describe the task of a classifier as trying to determine (or, predict) the value of some unknown variable  $y_i \in Y$  based on an observed variable  $x_i \in X$ . In the case of supervised learning, the ML algorithm is trained on a series of labelled data, taking the form  $(x_1, y_1), \dots, (x_n, y_n)$ , where each example is a pair  $(x_i, y_i)$  of an instance  $x_i$  and a label  $y_i$ . The goal is to learn an optimal mapping function (given certain pre-specified constraints) from the domain of possible values for  $X$  to the range of values that the target variable  $Y$  can assume. This formulation of the classification task covers many concrete examples and algorithm types, at a high level of abstraction (e.g. risk assessment, automated credit scoring, object identification).

<sup>13</sup> For the purpose of this illustration, we will not worry about the specific details of  $X$  or  $Y$ . However, the general format of this case study is similar to many widely used scoring systems, which need not rely on ML to function (e.g. [64]).



**Fig. 4** (Non-exhaustive) options for presenting an argument that is oriented towards multiple ethical goals. In this instance, the options present a case where the goal of explainability is subsumed within a higher-level goal of fairness (a), vice-versa (b), a case where both

goals are treated as jointly important and interlinked in virtue of the lower level claims they depend upon (c), and an option that builds two separate assurance cases (d)

phenotypic, or physiological characteristics), and (b) its potential to influence the healthcare professional’s diagnosis and subsequent recommendation for treatment.

However, the team also work with patients and other stakeholders during the initial project planning to determine how a goal related to greater health equity is understood by those affected or impacted by such a system. Following this analysis and dialogue, the team recognise that any attempt to provide assurance for a claim regarding greater health equity must also incorporate considerations of how their system supports a complementary principle (and goal) such as ‘explainability.’ They reach this understanding, for instance, on the basis of reflective dialogue with stakeholders about the importance of additional ethical considerations in healthcare, such as respect for patient autonomy and informed consent in decision-making.

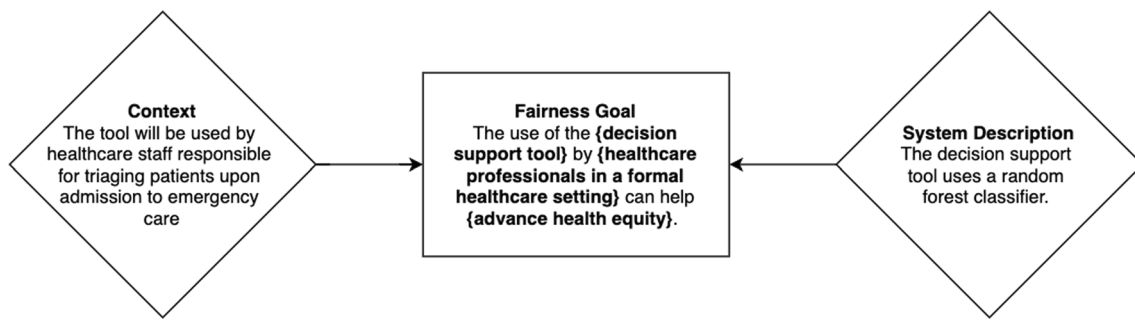
We make no specific claims about how these goals ought to relate in any given assurance case, as they are merely illustrative, and the specification of ethical principles requires careful consideration of context. However, Fig. 4 shows how the goals could, in principle, relate to each other. The options are (a) fairness could be the primary normative goal but depend upon the realisation of a sub-goal of explainability, perhaps due to the importance of informed consent; (b) vice versa; (c) both goals are jointly important but interlinked in virtue of the lower-level claims they depend upon; or (d) the two goals are independent and require two separate assurance cases. For simplicity, we will treat the goal of fairness as dependant on explainability for the purpose of our illustrative example. Moreover, we acknowledge that other values and principles may be involved (e.g. accountability), but choose to focus on fairness and explainability for simplicity of exposition.

Starting with the identification of relevant ethical values and principles supports the kind of anticipatory reflection that is emphasised by work in responsible research and innovation [66]. In ethical assurance, this reflective process is guided by the scaffolding of the project lifecycle (see Fig. 2).

For example, at the project planning stage, the project team may identify that they need to consider how relevant social determinants of health could affect the accuracy of the system for certain groups [49], or whether existing social biases that favour other sub-groups may lead to unfair levels of access to the system. If the system is intended for use in an area where a high percentage of patients are exposed to poor working conditions, an important reflective question to ask would be, ‘Does the model include relevant variables that can measure the relevant social risk factors?’ If the answer to this question is ‘no,’ the system may fail to accurately assess this sub-group. However, if a gap like this is identified early enough, perhaps as a result of engaging relevant experts or stakeholders during the problem formulation stage, or during exploratory data analysis, then it may be sufficient to alter plans for the data extraction and procurement stage that aim to improve the representativeness of the dataset. If so, the team could consequently select a relevant fairness optimisation constraint that promotes greater health equity throughout the target population, and which could be integrated during (model) development (e.g. during preprocessing, training, or post-processing), and then verified during model testing and validation and reported on during model reporting. Alternatively, the process of identifying the top-level goal may lead to a realisation that the function and purpose of the tool is poorly understood, or that the potential benefits are outweighed by the potential risks. If so, then the right decision may be to not proceed with the project at all.<sup>14</sup>

Therefore, using the project lifecycle as a guide at the outset of a project, the project team can reflect upon the possible decision points and challenges that are likely to

<sup>14</sup> This is just a selection of considerations. We cannot hope to cover all other relevant topics here, such as the importance of ensuring that the fairness optimisation constraints are considered reasonable by the affected stakeholders.



**Fig. 5** An example of how the top-level normative goal can be further contextualised and specified through supporting elements

emerge throughout the project, and which have a bearing on the selected goal.

In defining a top-level normative goal, it is not enough to simply state that the goal is ‘fairness’ or ‘explainability.’ This is because a top-level normative goal that simply stated, “This decision support tool is fair.” would be insufficiently specified, and would simply give rise to the question, ‘What notion of fairness is being employed?’ Therefore, it is important to acknowledge how the context of the project provides the basis for the specification of the principles or goal.

In the case of our hypothetical example, we can assume that a first pass at a normative goal would be something like the following:

“The use of the decision support tool helps advance health equity.”

However, this is missing several vital pieces of information, and can be made clearer by highlighting the relevant components of the goal. A better version would be:

“The use of the {decision support tool} by {healthcare professionals in a formal healthcare setting} can help advance {health equity}.”

Here, the brackets, respectively, highlight several important elements that feed into the top-level normative goal:

- a description of the {technological system},
- the {context} in which the system is being deployed, and
- the {normative goal} that centres the assurance case

This allows further details for the components to be given in the linked elements, as displayed in Fig. 5. For instance, providing a short description of the type of ML algorithm that is implemented within the {technological system} or providing further specification to help clarify the {normative goal} (e.g. a definition of the notion of ‘fairness’ or ‘health equity’ that is being employed). As the subsequent

components provide further detail that build on this initial goal, keeping the formulation of the top-level goal concise is advised.<sup>15</sup>

#### 4.2.2 Project and system property claims

Once the top-level normative goal has been sufficiently specified, it is then necessary to identify the actual properties of the project and system that help operationalise the set of ethical principles that define the goal of the project. This includes identifying the decisions and actions taken throughout the project’s lifecycle that ensure the goal is achieved (e.g. robust information governance processes to protect sensitive healthcare information).

In our running case study, the project team may have identified several (non-exhaustive) properties of their system or project that are relevant to their goal, using the project lifecycle as a guide. These properties can be formulated as statements about actions or decisions that need to be taken during specific stages:

- “During exploratory data analysis we must consider the possibility that diagnostic access bias<sup>16</sup> has affected the quality of our training data.”
- “At the model reporting stage, it will be important to ensure that information about the representativeness of our dataset is recorded, while also remaining sensitive to the need to maintain data privacy. Therefore, we will need to decide how granular to make our data while

<sup>15</sup> This also connects with some possible, future directions for ethical assurance that we discuss in §5.3. Specifically, the possibility of modularising ethical assurance to support the development of argument patterns or a model-based approach.

<sup>16</sup> Diagnostic access bias arises when individuals differ in their geographic, temporal, and economic access to healthcare services, this variation may result in their exclusion from a study or dataset, differential access to diagnostic tests, or affect the accuracy of the diagnostic test itself. This can cause under- or over-estimation of the true prevalence of a disease, and lead to worse treatment for socioeconomically deprived individuals.

remaining sensitive to any potential trade-offs with accuracy metrics among patient sub-groups.”

- “The healthcare professionals should be able to investigate and challenge the rationale for a particular assessment outcome during system use and monitoring, to ensure that professional judgement acts as a safeguard against false positives and false negatives, while also supporting explainability.”

There are, of course, many more decisions that are relevant to the top-level normative goal—our aim here is just to highlight some illustrative examples. However, the process of using the project lifecycle to support anticipatory reflection can help to uncover the properties of the project or system and identify possible actions that could be undertaken.

For example, let us say the project team decide to consult a group of clinical experts and social care workers with knowledge of the site from which the data were generated to determine the likelihood that the risk of diagnostic access bias has been minimised, as well as any other significant statistical, social, or cognitive biases. They could then formulate the following claim about a property of their project:

“We consulted a panel of experts to independently assess our dataset and ensure that the effect of bias has been minimised prior to model development.”

This claim may be a necessary component in developing a convincing assurance case but is insufficient on its own to justify the top-level goal for at least two reasons.

First, additional claims will be required to jointly satisfy the top-level goal. Some of these claims may be subordinate to the parent claim (e.g. claims about which types of bias were assessed). In addition, some of these claims (as above) refer to aspects of the project’s management (e.g. choices about how the project was managed), rather than to aspects of the system itself (e.g. details about the user interface of the decision support tool). Both sets (i.e. project and system property claims) reflect important sources of procedural claims that aim to legitimise the project’s overall governance.

#### 4.2.3 Evidential claim and artefact

Once the project or system property claims have been identified, it will be necessary to link the claims to documented evidence, or ‘evidential artefacts’ (e.g. a report that details the outputs from the model testing or validation stage).

Developing the project and system property claims, and gathering the necessary supporting evidence, will, in practice, be a simultaneous process—again, highlighting the iterative and cyclical nature of both the project lifecycle and the reflect, act, and justify process.

Whether the formulation of a system or project property claim requires evidence will, of course, depend on the nature of the claim. As Cartwright and Hardie ([13], 53) acknowledge,

“Some claims are self-evident or already well established. They do not need to be backed up by anything further for you to be justified in taking them to be true.”

In these cases, there may be no need to document an ‘evidential artefact.’ Whereas, in other cases an evidential artefact will need to be referenced and an ‘evidential claim’ will need to be established. An evidential claim can be treated as a proposition or description of the relevant evidence.

The intended audience of the assurance case itself will play a part in determining whether evidence is required, and how the evidential claim is formulated. For instance, if the party responsible for developing an assurance case is trusted by the party reviewing it, they may be willing to accept a propositional claim as evidence, rather than demanding further documented evidence (e.g. an external auditor versus an internal red team).

The evidence may also substantiate multiple claims if the system property claims and argument are wide in scope. As such, there may be a many-to-one relationship between evidential claims and an artefact. For example, consider the two system/project claims about our hypothetical project depicted in Fig. 6. As supporting evidence for these two claims, the assurance case may refer to the findings of an equality impact assessment undertaken at the outset of the project, which in this case may serve as a single evidential artefact that helps justify the two claims (with appropriate reference to specific sections of the assessment). This is important, as it means that certain pieces of documentation, which may be generated during typical project activities, such as algorithmic impact assessments [44, 62], transparency statements [18], or datasheets [26] could help ground several of the claims being made. In turn, regulators could incentivise or mandate certain activities that are important sources of evidence for ethical assurance cases.

In many cases, establishing a complete justificatory link between a system/project claim and an evidential claim requires one final step.

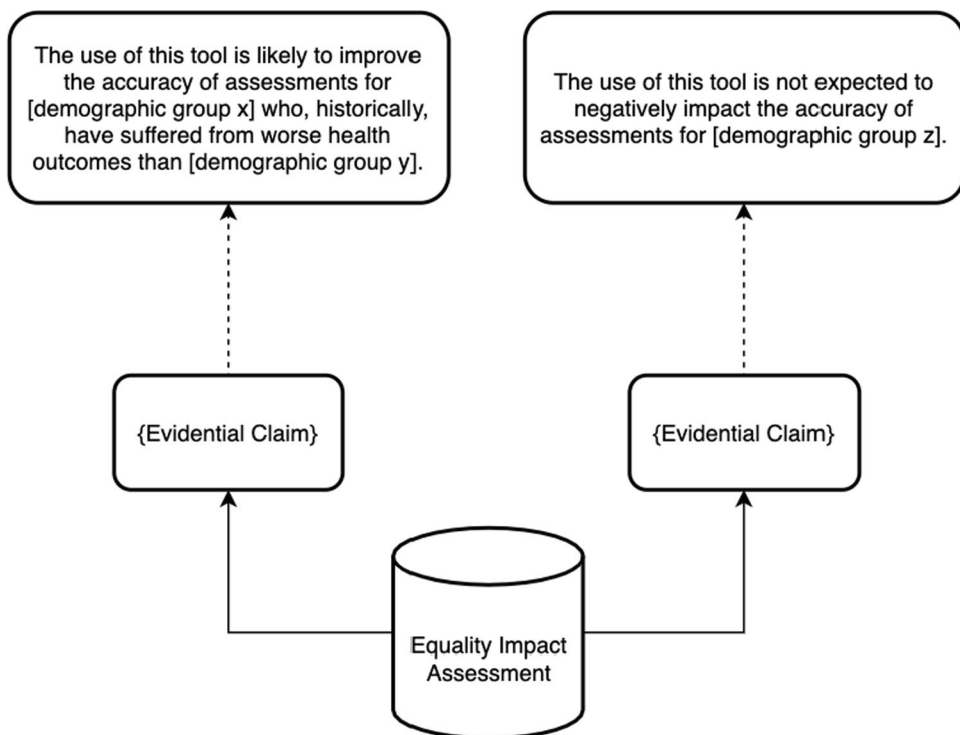
#### 4.2.4 Argument (warrant)

The final step in the practical process of building an ethical assurance case is more nuanced than the previous steps but is crucial for making key assumptions explicit and ensuring that the standpoint on which the argument depends is clearly demonstrated.

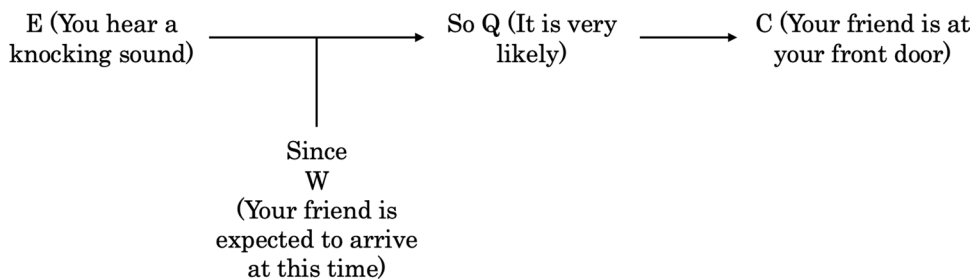
Consider the following two claims:



**Fig. 6** A portion of an assurance case showing how two claims about a project can be supported by the same evidential artefact



**Fig. 7** Diagram showing the relationship between several propositions: evidence (E), warrant (W), and a qualified claim (C)



1. System/project property claim: the {ML model} produces fair outputs because it has been trained with {fairness optimisation constraint x}.
2. Evidential claim: the results of the {fairness optimisation process} were {...}.

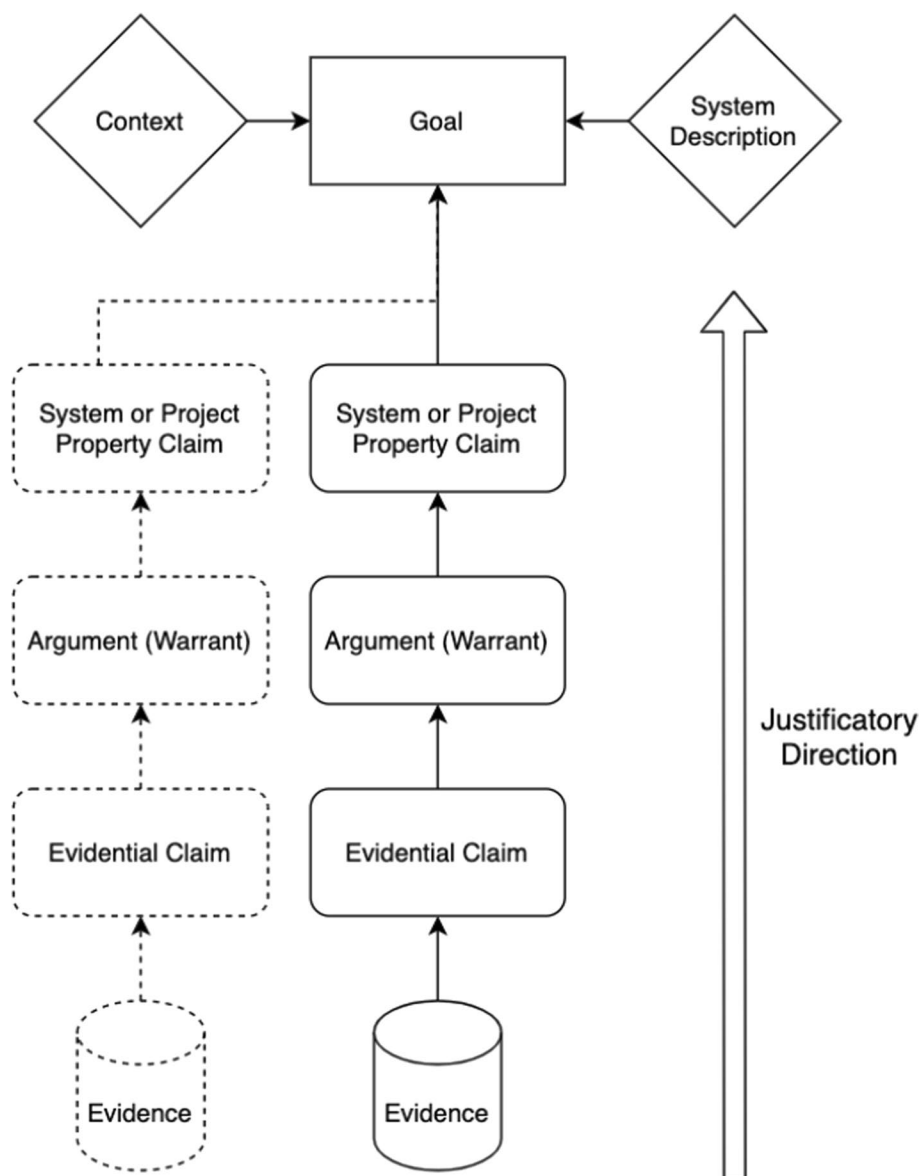
The role of the evidential claim here is intended to support the system/project property claim. However, whether an individual is justified in assenting to a belief in the system claim depends on a missing assumption between the two claims—whether the fairness optimisation process is reliable or appropriate.

The philosopher Stephen Toulmin [69], whose work heavily influenced the development of argument-based assurance, referred to this link as the ‘warrant.’ Simply put, the warrant is a step in an argument that links some evidence to a particular propositional claim, with the possible addition of a qualifier. This is indicated in Fig. 7 with a toy example.

In the above example, the inference from the evidence (You hear a knocking sound) to the qualified claim (it is very likely your friend is at your door) is only justified when we also provide additional warrant (your friend is expected to arrive at this time). Identifying the warrant that supports the inferential step from evidence to a claim can be one of the most challenging and complex steps in developing an assurance case, as it requires the project team to make explicit many of their assumptions.<sup>17</sup> For example, let us assume our hypothetical project team decide to adjust their classifier during post-processing so that it is uncorrelated with a protected attribute, and subsequently provide evidence about the model’s performance. To establish a link between the

<sup>17</sup> It also involves what epistemologists refer to as the transmission of justification across inference, which is a process where the justification for one belief (*p*) derives its justification from the justification that one has for a secondary belief (*q*) [52].

**Fig. 8** The general structure of an ethical assurance case, indicating the various elements and their relationship to each other



evidential claim about the model’s evaluation and a system claim relating to, say, group fairness, the project team will also need to add something like the following:

“The use of [fairness optimisation constraint  $x$ ] is legitimate in this context as it was selected by a representative group of stakeholders and experts, during a process of deliberative dialogue, as the most appropriate technique for operationalising and embedding our notion of fairness within the project.”

In practice, the project team may have to construct the full argument once the surrounding components (i.e. ‘system property claim’ and ‘evidential claim’) have all been selected, and the inferential links between the set of claims

is apparent. Although there is a concern that this may lead to the construction of post hoc arguments, it may also be necessary to fully evaluate whether the selected evidence is sufficient to support the argument (see Sect. 4.4.1 on determining and evaluating evidence). For instance, the project team may realise that a link between two claims rests on a faulty assumption, and that further work will be required. Again, the sooner that an anticipatory and ongoing process of reflection, action, and justification begins, the more likely such gaps will be identified and addressed.

Now that we have seen the practical direction for building an assurance case, we are in a better position to understand the complementary justificatory direction that links the elements together.

### 4.3 The justificatory direction

Figure 8 offers a representation of the logical structure that links each of the elements together. The directionality of the arrows represents the fact that the lower-level components provide inferential support for the higher level claims (e.g. an argument that warrants a system property claim; an evidential claim that raises the likelihood of an argument being valid). This is why the arrows lead upwards, or are oriented towards the top-level normative goal.<sup>18</sup> However, in practice, the directionality of constructing an assurance case, based on deliberative practices and actionable decisions, leads to a bidirectionality of the method of assurance, as it is likely that goals and claims will be delineated prior to evidence being accumulated.

In addition to this goal-oriented focus, ethical assurance also has a systems-level focus. Like traditional safety cases, the complexity of an ethical assurance case will depend on the complexity of the system itself (and the social context in which it is deployed), and the level of detail required will be guided to some degree by a principle of proportionality. Ethical assurance departs from tools like GSN (see Sect. 2), however, in that it aims for representational simplicity rather than completeness. As such, it has a lot less representational expressivity. This is a purposeful design consideration to improve accessibility for the purpose of supporting inclusive, participatory methods.

### 4.4 Building and using an ethical assurance case

In closing this section, we expand on three additional topics regarding the development of an ethical assurance case:

1. Determining and evaluating evidence (i.e. how the evidential artefacts that ground an assurance case are selected by the project team).
2. Phased assurance and active enquiry (i.e. how an assurance case can be built iteratively throughout the project lifecycle through active participation with stakeholders).
3. Argument patterns (i.e. how generalisable templates can be developed to support best practices across and within domains, such as healthcare or education).

<sup>18</sup> Those readers who are familiar with informal logic and argumentation theory will recognise that this structure is also heavily influenced by the work of Stephen Toulmin [69], whose research into the structure of arguments has been highly influential in the development of ABA.

#### 4.4.1 Determining and evaluating evidence

As we have demonstrated, the process of developing an ethical assurance case relies on the collection and use of evidence. But how does one go about evaluating the probative value of evidence in the context of ethical assurance? In addition, what counts as evidence in the first place? Drawing upon concepts from argumentation theory and jurisprudence can help answer these questions, especially given our context of argument-based assurance.

Recall that what is being developed here is a structured case that can be used to support a process of reason-giving (on the part of the project team). This act of giving reasons implies a recipient and an active process of dialogue—even if it is asynchronous and mediated through a formal process, rather than a verbal one. Although the recipient(s) of an assurance case will vary, we can treat the exchange of reasons, structured as an assurance case, as similar to the determination, use, and interrogation of evidence in a legal setting (e.g. arguing a case to a judge or jury, which is critiqued by another party).

Three principles guide fact finding in a legal case: relevance, materiality, and admissibility (see [35] for a more detailed exposition). ‘Relevance’ is a relational concept that holds between two propositions. In the present context, this is the relevance of the evidence in establishing (qualified) warrant for the respective claim. ‘Materiality’ refers to whether a fact is receivable by a court, which in turn depends on additional legal facts. For instance, a fact that is not in dispute may be relevant in the ordinary sense of the term, but not material to a court’s decision (e.g. a fact about a breach of contract that is accepted by both parties). The materiality of evidence is significant because it reduces the need to consider (relevant) evidence that will not affect the outcome of the (assurance) case. ‘Admissibility’ is based on the rule of law and covers conditions for exclusions that go beyond relevance or materiality. For instance, a fact may be both relevant and material but still be inadmissible due to additional legal rules (e.g. hearsay is not admissible even if it is relevant and material).

Like the question of what constitutes relevant, material, and admissible evidence in a court, we can ask ‘under what conditions can an evidential claim or documented evidence serve as grounds for reasonable inference in an assurance case?’ Here, we are aiming to assess the probative value of a particular evidential artefact (or associated claim) within an assurance case. For example, anecdotal evidence may be accepted in everyday conversation but would not be admitted as evidence in court.

Furthermore, when evaluating the probative value of evidence within an assurance case, we must do so with an eye to the whole case, not just the isolated value of a particular claim or artefact. This requires consideration of:

1. The probative value of each evidential artefact/claim in relation to a specific project/system claim
2. The sufficiency of either an individual evidential artefact/claim or set of artefacts/claims in relation to a specific project/system claim
3. The sufficiency of the overall assurance case conditional on the top-level normative goal<sup>19</sup>

There is a further sense in which the evaluation of evidence is relational. Whether a propositional claim or documented artefact counts as evidence depends on whether it is judged by the recipient as supporting a process of reasonable inference. That is, in a process of dialogical reasoning, does the evidence serve as a reason in support of the conclusion that a proposition (i.e. the claim) is true or false, probable or improbable, for the parties involved? Here, it is important that the top-level normative goal is grounded in a value or principle (e.g. human dignity) that is accepted as a shared and public grounds for communication. Hence, the need to base an argument on the shared acceptance of reasonable ethical values and principles [44].

There is, understandably, no general answer that can be offered here when it comes to the determination and evaluation of specific evidential artefacts and claims. However, raising this question nevertheless clarifies a previously mentioned function of ethical assurance that is sometimes neglected in AI/ML assurance: the need to support active enquiry.

#### 4.4.2 Phased assurance and active enquiry

It is already well recognised that the method of ABA goes beyond the mere documentation of development processes to support compliance and regulatory obligations [15]. Instead, the development of an assurance case can, among other things, assist anticipatory reflection and deliberation (as seen above), and help build and establish trust through transparent communication.

As an ethical standard for algorithmic accountability, however, the goal of transparency has been criticised on several grounds. For instance, some researchers have pointed to legal and financial barriers to achieving transparency, resulting in the need to develop and employ techniques from areas such as computational journalism to get around these barriers and assess the use of algorithmic systems in high-stake domains (e.g. advertising, criminal justice) [21]. Unfortunately, these techniques can be limited in significant ways and transparency is not equivalent to governance or control [22]. Other researchers have questioned the ideal

of transparency more generally, arguing that it can be used as a means to intentionally occlude, as with the case of a company that provides mountains of evidence to be sifted through at the last minute prior to a legal case [1]. To help address these concerns about transparency, we can turn to the work of the moral philosopher Onora O’Neill.

In her BBC Reith Lectures, *A Question of Trust*, O’Neill [55, 77–78] argues that,

“[...] if we want a society in which placing trust is feasible we need to look for ways in which we can *actively check* one another’s claims.”

She also acknowledges,

“[...] active checking of information is pretty hard for many of us. Unqualified trust is then understandably rather scarce.”

This notion of ‘active checking’ is important. In the context of ethical assurance, we can think of it as a form of dialogical communication that we label ‘active enquiry.’ That is, the development of an ethical assurance case ought not be approached as a monological exercise, in which a project team produce an argument that is simply presented as fact to relevant stakeholders. Rather, each argument is necessarily defeasible, because of both its inferential structure and the possibility that the acceptability or weighting of the normative goal may be legitimately challenged by certain stakeholders. This contestability of an assurance case means it should be seen as a living document that can form the basis for an ongoing conversation about the acceptability of sociotechnical systems in key areas of society.

This connects to the second point. The iterative, ongoing, and situated nature of the design, development, and deployment of data-driven technologies requires a phased approach to ABA. Kelly [41] refers to this as the presentation of an “evolving safety argument,” outlining three stages: a preliminary stage, interim stage, and operational stage.

In the first stage, only an outline of the argument is produced, showing the principal objectives, and anticipated evidence. In the second stage, the argument is developed to reflect the increased knowledge of the design and specification of the system. Finally, the case evolves to reflect evidence that concerns how the system is tested and implemented. This general strategy is embedded in the Goal Structuring Notation (GSN) approach, but its influence should also be clear in the reflect, act, and justify approach to ethical assurance (see Fig. 3). However, we must take Kelly’s evolving safety argument approach one step further, and join the two ends together, to capture the cyclical and socially situated nature of ethical assurance. In doing so, the evolving nature of phased assurance aligns with the idea that moral deliberation and public ethical reasoning is best understood as a process of reflective equilibrium [61].

<sup>19</sup> The sufficiency of the overall assurance case will, of course, depend on 1 and 2.

As social norms and practices evolve, so too do the ethical demands and expectations on sociotechnical innovation. Exercising responsibility and maintaining trust, therefore, requires an inclusive, deliberative approach that remains responsive and open to new moral horizons—hence the circularity in Fig. 2.

The third and final point that emerges from O’Neill’s analysis is captured by the recognition that active enquiry is often limited by the aforementioned epistemic barriers such as intellectual property rights that restrict access to information, or limited technical literacy or capacity that inhibits an individual’s understanding of complex technical systems. Moreover, as organisations start exploring automated ways of testing, validating, and documenting ML development—such as Google’s automation of model reporting activities [25]—there is a very real risk that developers become more and more epistemically detached from the project lifecycle. This could, in turn, reduce opportunities for vital ethical reflection, such that assurance becomes an automated process that churns out floods of documentation—this would represent a form of “unintelligent accountability,” to appropriate a term from O’Neill [55]. Returning to the raised concern about transparency, this would place a disproportionate burden on individual users or groups of stakeholders to “seek out information about a system, to interpret that information, and determine its significance” (2018, 979). To counteract this possible trend, we need to continue to consider which practical mechanisms best support ethical assurance in general and processes of active enquiry and phased assurance more specifically. There are many open questions about how this can be achieved (see Sect. 5). In the next section, we suggest one mechanism of direct relevance to ethical assurance.

#### 4.4.3 (Ethical) argument patterns

Argument patterns are reusable templates for assurance claims, which address the types of claims, evidence, warrant and additional contextual information that must be covered to justify a claim pertaining to a particular normative goal. Is it possible to develop ethical argument patterns? If so, what role would they play?

Let us start with the first question. How could we develop ethical argument patterns? Although patterns could be proposed in a prescriptive, top-down manner, in the case of ethical assurance they are likely to have greater normative force and legitimacy if they arise bottom-up from actual patterns that are identifiable and generalisable from existing assurance cases. For instance, if several ethical assurance cases that were concerned with the goal of ‘protecting human dignity’ when deploying automated decision-making systems in criminal justice, all converged on a shared argument structure, then, *prima facie*, this would be reasonable

grounds for inferring that the structure represented a reliable argumentation pattern. As such, the pattern could be used as a starting point for subsequent assurance cases, and some of the elements could perhaps take on the role of so-called “default reasons” [37].

An alternative means by which patterns could be developed is through stakeholder engagement activities. For instance, activities including deliberative dialogue [2], in which stakeholders consider relevant information from multiple points of view, enables the exploration and discussion of topics or issues, without assuming the prior existence of consensus or agreement. Instead, the process is designed to build consensus such that any judgements on the topics or issues under discussion are arrived at through careful consideration and a greater understanding and awareness of possible tensions. These types of engagement activities could support the development of argument patterns, by exploring issues such as socially desirable ethical goals, acceptable means of specification, convincing and persuasive arguments, and accessible forms of evidence.

Working with stakeholder groups in this way would also help expand the role of both ethical assurance and ethical argument patterns. First, deliberative dialogue is not simply a method of opinion polling or aggregation [23]. More than this, the focus on dialogue and consensus building helps orient the activities towards a more dialogical mode of education, which could help improve digital literacy for the participants. Second, by engaging stakeholders in the assurance process, they are more likely to have trust in the systems being deployed, as they will either recognise their own values in the assurance case/patterns, or at least have a greater understanding of the tensions and trade-offs that have been reflected upon throughout a project’s lifecycle.

The three topics we have explored—determining and evaluating evidence, phased assurance and active enquiry, and ethical argument patterns—help demonstrate the potential value of ethical assurance. We now turn to consider some possible challenges, open questions, and next steps.

## 5 Conclusion: challenges, open questions, and next steps

The proposals we have offered in this article are still very much in their nascency, but we believe that ethical assurance has a lot of promise for supporting and complementing ongoing attempts to ensure that data-driven technologies work to promote individual and social well-being.

To further strengthen the support for our methodology, and in closing this paper, we anticipate some potential challenges for ethical assurance and offer some responses. We also note the need for further research and suggest next steps, to which we hope others will choose to contribute, to

help turn this proposal into an active programme of research and development.

## 5.1 Challenges

### 5.1.1 Ethical assurance could be misused

This worry is reflected in claims that AI ethics supports cases of “ethics-washing” [32] or as PR coverage for ethically problematic institutional practices. The worry can, of course, be extended to ethical assurance (e.g. ethical assurance could be used as a mean for legitimising or covering up unethical projects).

An immediate response is to simply note that this is a risk for nearly all practical mechanisms that operate in this space, as tools can be used for good and for bad. However, this response is unsatisfying. A better response would be to note that the worry is itself mitigated by the way in which ethical assurance has been designed.

By exposing the argumentation structure to open critique and active enquiry, an ethical assurance case is more likely to expose an unconvincing or incomplete argument. In turn, there is a potential for improving the argument, or using available legal mechanisms to hold the organisation accountable—an improvement on the limited notions of transparency noted above. Moreover, the reflective and anticipatory approach to the design, development, and deployment of algorithmic systems that is enjoined by ethical assurance prevents the superficiality of “ethics washing” practices in virtue of the requirement that documented bridges be built between actions and justifications across the entire AI/ML lifecycle.

### 5.1.2 Ethical assurance is too demanding

No one said ethics was easy. Doing the right thing can be challenging and time-consuming, and in some cases, the costs of failing to consider ethics can be greater than the costs of doing the right thing in the first place. Nevertheless, this is a valid concern when we consider the limited capacity, available resources, and skills that often inhibit smaller businesses and public sector organisations. Therefore, the challenge here is to address how smaller organisations will manage the increased demand that ethical assurance places upon them, alongside existing regulatory requirements.

First, ethical assurance should not be misconstrued as a form of compliance. It is not a separate requirement akin to a data protection impact assessment, but rather a scaffold to support and supplement these pre-existing requirements while promoting virtues such as transparent communication and public or stakeholder accountability. While an ethical assurance case may take time to produce, the overarching process has been designed to complement

and extend existing regulatory requirements and emerging best practices of design, development, and deployment, rather than to be an additional burden to be completed at the end of a project. This is why the project lifecycle is important as an anticipatory guide.

However, it is still important to keep in mind a principle of proportionality. Some projects are exposed to greater risk due to the context in which they are developed (e.g. healthcare). Where a project has very low risk, it may not be necessary to develop an extensive ethical assurance case, but instead just use the methodology to guide a process of reflection and document this accordingly. Here, argument patterns may serve a further role by helping developers assess and evaluate possible risks or benefits during the project design stage.

In addition, ethical assurance will be made less demanding by improving the skills and capacities of regulators and auditors, perhaps exploring ethical standards and certification schemes that complement ethical assurance (e.g. as forms of trusted evidence) and make the process more efficient. Ultimately, ethical assurance is a framework that can be made increasingly easy to employ as additional supporting mechanisms emerge.

### 5.1.3 Ethical assurance cases are defeasible

This challenge focuses on the inferential support that the evidence provides for both the property/system claims and the top-level normative goal. As such, the overall argument will be defeasible, remaining open in principle to revision, based on objections or competing forms of evidence. This may seem, *prima facie*, like a problem for the methodology. However, on deeper reflection it becomes clear that this is a feature (and not a bug), which supports the function and purpose of assurance.

As we discussed in Sects. 3 and 4, ethical assurance is designed to support a process of reflection, action, and justification throughout a project’s lifecycle, and also enable the active enquiry of stakeholders both prior to and after a system is implemented within a particular context. For many technologies, their social impact will not be made apparent until the time of deployment. A method of assurance that failed to acknowledge these uncertainties would fail to exercise an appropriate level of epistemic humility. In contrast, the defeasibility of an ethical argument can (a) help motivate the need for an inclusive and participatory approach to design, development, and deployment, (b) encourage developers to adopt a phased and modular approach to building an assurance case, and (c) ensure that possible visions for ethically and socially desirable futures remain open to consideration.

## 5.2 Open questions

### 5.2.1 How should developers deal with sensitive or probabilistic evidence?

This is in fact two questions, both of which pertain to the use of evidence: How should sensitive evidence that cannot be made public be incorporated into an assurance case? How should probabilistic evidence be used and evaluated?

In the case of sensitive evidence, it will be important to develop the capacity of independent auditors, to ensure appropriate levels of transparency. Likewise, it will be important for producers of ethical assurance cases to develop layered or tiered methods of presenting their cases, so that sensitive information can differentially be made available to relevant parties and the that needs of non-technical stakeholders can be accommodated by a plain language and easily understandable layer of presentation [39] (also see Sect. 5.3.2).

However, for both questions, we must also encourage the development and adoption of norms and best practices within and across domains and industries. For instance, in the case of an autonomous vehicle, current efforts to deal with probabilistic evidence to support safety claims (e.g. this vehicle has not caused an injury in  $x$  number of miles or journeys) are based on the notion of reducing risk to levels that are as low as reasonably practicable (known as the ALARP principle) [19]. Industry norms play a role in determining what is “reasonably practicable,” but it is not presently clear whether or how this risk-based approach will transfer to the case of ethical assurance.

### 5.2.2 How does ethical assurance work for projects that distribute responsibilities across teams and organisations?

The stages of design, development, and deployment for a complex ML-based system, and the tasks within these stages, may be carried out by different teams, across multiple organisations. For instance, the increasing availability of so-called “model libraries”—repositories for pre-trained models—means that the procurement of components may go beyond that of just the data or other services necessary for deploying a system. As such, the evidence required to fulfil an ethical assurance case may require the coordination of multiple actors.

Research into modular assurance cases remains an open question even in the comparatively well-established safety case literature [29]. However, there is good reason to think that, in conjunction with a phased approach to development, a modular form of ethical assurance cases could support distributed project management. In addition, assurance contracts, which hold parties legally accountable for the claims

made within their “module” (e.g. that they are not falsifying evidence) could also be explored to support more complex cases and maintain trust.

## 5.3 Next steps

### 5.3.1 Evaluating efficacy of ethical assurance

Ethical assurance is a type of argument-based assurance, and, therefore, has many of the same benefits of ABA that were outlined in §2. However, the extent to which these benefits accrue to projects that employ assurance cases is a matter that remains inconclusive [29]. For instance, to what extent do safety cases contribute to the safety-related outcomes of a project? Those who question the efficacy of safety cases can point to well-known failures, such as the loss of the RAF Nimrod MR2 Aircraft XV230 in Afghanistan in 2006, in support of their scepticism. In an independent review of this incident, the safety case that was drawn up for the aircraft by BAE systems was described as “a lamentable job from start to finish. It was riddled with errors. It missed the key dangers. Its production is a story of incompetence, complacency, and cynicism” ([31], 10). Findings such as these do not appear to paint a positive picture for the efficacy of assurance cases. However, it would be premature to write ABA off altogether.

It is clear from the Nimrod review that part of the failure stems from organisational malaise among the producer of the safety case. As the review notes, “the task of drawing up the Safety Case became essentially a paperwork and ‘tick-box’ exercise” ([31], 10). In earlier sections, we have been careful to acknowledge that the production of an assurance case is only one component in a broader commitment to ethical reflection, deliberation and participatory design, development, and deployment. Such processes can only flourish in organisations and teams that take care to build a culture of readiness, reflection, and responsible research and innovation. This was evidently not the case with the production of the Nimrod safety case, as the review goes on to acknowledge that the above matters raised “question marks about the prevailing ethical culture at BAE Systems” ([31], 10). Others have also recognised the importance of organisational culture—construed broadly as both interpersonal norms and values (e.g. culture of reflection), as well as external constraints or practical mechanisms (e.g. incident reporting systems). For example, in discussing the application of safety cases to healthcare, Sujan and Habli ([67], 4) are optimistic about the possibility of using safety cases for improving digital health innovations, but only when accompanied by “far-reaching structural changes.”

In addition to recognising the importance of organisational culture, Sujan and Habli [67] also offer two reasons why there is no conclusive evidence for the efficacy of safety

assurance more generally. First, safety cases are typically employed as regulatory instruments in “high-hazard settings,” characterised by high-severity but low-frequency events (e.g. the catastrophic failure and loss of an aircraft, as above). As such, it is difficult to obtain statistically meaningful data about the casual impact that safety cases have. Second, the process of developing a safety case varies widely across domains (e.g. nuclear, aviation), making it hard to determine which factors of a safety case or the supporting culture are causally relevant to increasing or decreasing safety risks.

While these challenges have been raised in connection with safety cases specifically, it is reasonable to conclude that similar challenges will apply to the possible adoption and uptake of ethical assurance cases. For example, in their critique of the ideal of transparency as it applies to algorithmic systems, Annany and Crawford ([1], 980) question the assumption that transparency engenders trust in organisations and systems, given the lack of confirmatory empirical research. As a next step, it will be necessary to identify and determine how different substantive and structural factors may contribute to the success or failure of ethical assurance.

### 5.3.2 Systems and standards

In addition to the directions implied by the previous two open questions, there is a further avenue that would help develop the current proposals into a more robust research programme: practical systems and standards that can help teams and organisations implement ethical assurance.

Ethical assurance cases could end up being large and complex, especially when multiple goals are interlinked (see Sect. 4.2.1). Therefore, it would be sensible to consider—as happens in ABA more generally—how software tools can support the process of both building and interacting with an ethical assurance case. For example, could online software platforms help teams iteratively construct an ethical assurance case and open it up to different groups of stakeholders?

To avoid the situation where we end up with a miscellany of tools and approaches, much like the problem we described at the start of this article, it would be wise to consider whether there are pre-existing standards and best practices to draw from (e.g. interoperability standards). For ABA more generally, the Object Management Group (OMG)—a computer industry standards consortium—has developed the Structured Assurance Case Metamodel (SACM), which is used to represent assurance case [56]. It also provides further constraints on the use of, say, argument patterns and relevant evidence, or acceptable syntax and semantics. As Hawkins et al. [33] acknowledge, having a (meta)model of the assurance process can bring certain benefits of model-driven approaches to engineering, such as automation, traceability and accountability,

transformation and validation. The value of this approach, however, is conditional on the extent to which it is adopted and supported by the community (e.g. used by developers, recognised by regulators). While we have avoided the issue of whether ethical assurance could conform to such a standard, there appears to be, *prima facie*, no reason why a model-based approach could not be pursued, to build, for instance, a metamodel that synthesises the reflect, act, and justify model (see Fig. 3) with the project lifecycle model (see Fig. 2) and informal logic of an ethical assurance case (see Fig. 8).

Pursuing such an approach could also increase the value of possible online software tools. For example, consider a metamodel that provided guidance on the adoption of argument patterns for ethical assurance cases that made use of multiple top-level normative goals. If particular goals (and supporting claims and evidence) were of more interest to a specific group of stakeholders (e.g. auditors), then the metamodel could enable the use of simple software filters that could be applied when interacting with an ethical assurance case, allowing the stakeholder group to focus in on only those parts of the argument that were relevant. Or, it could support the reframing of an argument at a different level of complexity when certain evidential claims (e.g. details about an ML component) require technical expertise that go beyond the scope of a certain group of stakeholders.

These capabilities would also depend on the development of ethical argument patterns. However, by developing a series of reusable templates (i.e. patterns) that have proved to be helpful in supporting critical reflection and deliberation in specific contexts or use cases, the development of ethical assurance cases could be made more efficient and effective. For instance, it would be possible to develop software tools linked to a structured repository of argument patterns (e.g. through an API), which prompt users with a series of questions about which concept of ‘fairness’ they are using, and offer guides for specifying high-level ethical principles in a specific context.

Therefore, an important next step here is to consider the formal and syntactical representation of ethical assurance in more detail. For present purposes, we have sidestepped issues such as which notation schema it will rely upon (e.g. GSN) and whether it will conform to existing standards (e.g. SACM) to focus on the broader purpose and motivation of the ethical assurance methodology. This gap will, of course, need to be addressed to ensure the full potential of ethical assurance is realised. For the time being, however, it is sufficient to conclude by noting that there are good reasons to believe that the development of systems and standards is a worthwhile avenue to explore. We hope that by outlining both the potential value of ethical assurance in general, and outlining concrete opportunities for its development, that an active community will emerge to help realise its potential for



ensuring that data-science and AI contribute to an inclusive and collective social good.

**Acknowledgements** We wish to thank Ibrahim Habli, Zoe Porter, Geoff Keeling, Rosamund Powell, and Mike Katell for their insightful comments on earlier drafts of this article, as well as offering suggestions for further research that took the article in valuable directions, which it otherwise would not have explored.

**Funding** This research was supported by a grant from the UKRI Trustworthy Autonomous Systems Hub, awarded to Dr Christopher Burr. Additional funding was provided by Engineering and Physical Sciences Research Council (EPSRC Grant# EP/T001569/1, EPSRC Grant# EP/W006022/1), Economic and Social Research Council (ESRC Grant # ES/T007354/1).

## Declarations

**Conflict of interest** On behalf of all the authors, the corresponding author states that there is no conflict of interest.

## References

- Ananny, M., Crawford, K.: Seeing without knowing: limitations of the transparency ideal and its application to algorithmic accountability. *New Media Soc* **20**(3), 973–989 (2018). <https://doi.org/10.1177/1461444816676645>
- Andersson, E., McLean, S., Parlak, M., Melvin, G.: From fairy tale to Reality: Dispelling the myths around citizen engagement. *Involve and the RSA* (2013)
- Arnold, M., Bellamy, R.K.E., Hind, M., Houde, S., Mehta, S., Mojsilović, A., Nair, R., et al.: FactSheets: increasing trust in AI services through supplier's declarations of conformity. *IBM J Res Dev* **63**(4/5), 1–13 (2019). <https://doi.org/10.1147/JRD.2019.2942288>
- Ashmore, R., Calinescu, R., Paterson, C.: Assuring the machine learning lifecycle: desiderata, methods, and challenges. [Cs, Stat], May (2019). <http://arxiv.org/abs/1905.04223>.
- Beauchamp, T.L., DeGrazia, D.: Principles and principlism. In: Khushf, G. (ed.) *Handbook of Bioethics*, pp. 55–74. Springer, Dordrecht (2004). [https://doi.org/10.1007/1-4020-2127-5\\_3](https://doi.org/10.1007/1-4020-2127-5_3)
- Beauchamp, T.L., Childress, J.F.: *Principles of Biomedical Ethics*, 7th edn. Oxford University Press, New York (2013)
- Bender, E.M., Friedman, B.: Data statements for natural language processing: toward mitigating system bias and enabling better science. *Trans Assoc Comput Linguist* **6**, 587–604 (2018). [https://doi.org/10.1162/tacl\\_a\\_00041](https://doi.org/10.1162/tacl_a_00041)
- Benjamin, R.: *Race After Technology: Abolitionist Tools for the New Jim Code*. Polity, Medford (2019)
- Binns, R.: What can political philosophy teach us about algorithmic fairness? *IEEE Secur. Privacy* **16**(3), 73–80 (2018). <https://doi.org/10.1109/MSP.2018.2701147>
- Bloomfield, R., Bishop, P.: Safety and assurance cases: past, present and possible future an adelard perspective. In: Dale, C., Anderson, T. (eds.) *Making Systems Safer*, pp. 51–67. Springer, London (2010). [https://doi.org/10.1007/978-1-84996-086-1\\_4](https://doi.org/10.1007/978-1-84996-086-1_4)
- Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., Khlaaf, H., et al.: Toward trustworthy AI development: mechanisms for supporting verifiable claims (2020). [arXiv:2004.07213](https://arxiv.org/abs/2004.07213) [Cs], <http://arxiv.org/abs/2004.07213>
- Burton, S., Habli, I., Lawton, T., McDermid, J., Morgan, P., Porter, Z.: Mind the gaps: assuring the safety of autonomous systems from an engineering, ethical, and legal perspective. *Artif. Intell.* **279**(February), 103201 (2020). <https://doi.org/10.1016/j.artint.2019.103201>
- Cartwright, N., Hardie, J.: *Evidence-based policy: a practical guide to doing it better*. Oxford University Press, Oxford (2012)
- CDEI.: *The Roadmap to an Effective AI Ecosystem*. Centre for Data Ethics and Innovation. [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/1039146/The\\_roadmap\\_to\\_an\\_effective\\_AI\\_assurance\\_ecosystem.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1039146/The_roadmap_to_an_effective_AI_assurance_ecosystem.pdf) (2021)
- Cleland, G.M., Habli, I., Medhurst, J., Health Foundation (Great Britain): Evidence: using safety cases in industry and healthcare (2012)
- Cobbe, J., Lee, M.S.A., Singh, J.: Reviewable automated decision-making: a framework for accountable algorithmic systems. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Virtual Event Canada, pp. 598–609 (2021) <https://doi.org/10.1145/3442188.3445921>.
- Collingridge, D.: *The Social Control of Technology*. St. Martin's Press, New York (1980)
- Collins, G.S., Reitsma, J.B., Altman, D.G., Moons, K.G.M.: Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Ann. Intern. Med.* **162**(1), 55 (2015). <https://doi.org/10.7326/M14-0697>
- Commission Law.: *Automated Vehicles: Summary of Consultation Paper 3 A Regulatory Framework for Automated Vehicles* (2020)
- Community GSN.: *GSN Community Standard (Version 2)*. The Assurance Case Working Group (2018)
- Diakopoulos, N.: Algorithmic accountability reporting: on the investigation of black boxes. *Tow Center for Digital Journalism* (2014)
- Diakopoulos, N.: Algorithmic accountability: journalistic investigation of computational power structures. *Digit. J.* **3**(3), 398–415 (2015). <https://doi.org/10.1080/21670811.2014.976411>
- Dryzek, J.S., List, C.: Social choice theory and deliberative democracy: a reconciliation. *Br. J. Political Sci.* **33**(1), 1–28 (2003)
- Van Eemeren, F.H., Grootendorst, R.: *A Systematic Theory of Argumentation: The Pragma-Dialectical Approach*. Cambridge University Press, Cambridge (2004)
- Fang, H., Miao, H.: Introducing the model card toolkit for easier model transparency reporting. *Google AI Blog* (2020)
- Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Daumé III, H., Crawford, K.: Datasheets for datasets. In: *Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning* (2018). <http://arxiv.org/abs/1803.09010>
- Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, W., Wallach, H., Daumé, III H., Crawford, K.: Datasheets for Datasets (2019). [arXiv:1803.09010](https://arxiv.org/abs/1803.09010) [Cs]. <http://arxiv.org/abs/1803.09010>.
- Habermas, J.: *On the Pragmatics of Communication*. MIT Press, Cambridge (1998)
- Habli, I., Alexander, R., Hawkins, R.: Safety cases: an impending crisis? In: *Safety-Critical Systems Symposium (SSS'21)*, 18 (2021)
- Habli, I., Alexander, R., Hawkins, R., Sujun, M., McDermid, J., Picardi, C., Lawton, T.: Enhancing COVID-19 decision making by creating an assurance case for epidemiological models. *BMJ Health Care Inform* **27**(3), e100165 (2020). <https://doi.org/10.1136/bmjhci-2020-100165>
- Haddon-Cave, C., Great Britain, Parliament, and House of Commons.: *The NIMROD Review: an independent review into the*

- broader issues surrounding the loss of the RAF Nimrod Mr2 Aircraft Xv230 in Afghanistan in 2006. Stationery Office, London (2009)
32. Hao, K.: In 2020, Let's Stop AI Ethics-Washing and Actually Do Something. MIT Technology Review (2019). <https://www.technologyreview.com/2019/12/27/57/ai-ethics-washing-time-to-act/>.
  33. Hawkins, R., Habli, I., Kolovos, D., Paige, R., Kelly, T.: Weaving an assurance case from design: a model-based approach. In: 2015 IEEE 16th international symposium on high assurance systems engineering. IEEE, Daytona Beach Shores, pp. 110–117 (2015) <https://doi.org/10.1109/HASE.2015.25>.
  34. Hawkins, R., Paterson, C., Picardi, C., Jia, Y., Calinescu, R., Habli, I.: Guidance on the Assurance of Machine Learning in Autonomous Systems." University of York: Assuring Autonomy International Programme (AAIP) (2021).
  35. Ho, H.L.: The legal concept of evidence. In: Edward, N.Z. (Ed.) The Stanford Encyclopedia of Philosophy, Winter 2015. Metaphysics Research Lab, Stanford University.
  36. Holland, S., Hosny, A., Newman, S., Joseph, J., Chmielinski, K.: The dataset nutrition label: a framework to drive higher data quality standards (2018).
  37. Horty, J.F.: Reasons as Defaults. Oxford University Press, New York (2014)
  38. ICO.: Guidance on the AI Auditing Framework. Information Commissioner's Office (2020)
  39. ICO, and Alan Turing Institute.: Explaining Decisions Made with AI (2020)
  40. Kalluri, P.: Don't ask if artificial intelligence is good or fair, ask how it shifts power. *Nature* **583**(7815), 169–269 (2020). <https://doi.org/10.1038/d41586-020-02003-2>
  41. Kelly, T.P. Arguing safety A systematic approach to managing safety cases. Ph.D. thesis, Department of Computer Science: University of York (1998).
  42. Kind, C.: The Term 'Ethical AI' Is Finally Starting to Mean Something | VentureBeat. *VentureBeat* (2020). <https://venturebeat.com/2020/08/23/the-term-ethical-ai-is-finally-starting-to-mean-something/>. Accessed 6 May 2021
  43. Kroll, J.A.: Outlining traceability: a principle for operationalizing accountability in computing systems. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. ACM, Virtual Event Canada, pp. 758–771 (2021) <https://doi.org/10.1145/3442188.3445937>.
  44. Leslie, D.: Understanding artificial intelligence ethics and safety. The Alan Turing Institute, London (2019)
  45. Leslie, D.: The Secret Life of Algorithms in the Time of COVID-19. The Alan Turing Institute (2020) <https://www.turing.ac.uk/blog/secret-life-algorithms-time-covid-19>.
  46. Leslie, D.: The arc of the data scientific universe. *Harvard Data Sci Rev* (2021). <https://doi.org/10.1162/99608f92.938a18d7>
  47. Leslie, D., Rincon, C., Burr, C., Aitken, Katell, M., & Briggs, M.: AI Sustainability in Practice: Part I. The Alan Turing Institute and the UK Office for AI (2022a)
  48. Leslie, D., Rincon, C., Burr, C., Aitken, Katell, M., & Briggs, M. (2022b). AI Sustainability in Practice: Part II. The Alan Turing Institute and the UK Office for AI
  49. Lucyk, K., McLaren, L.: Taking Stock of the Social Determinants of Health: a scoping review. Edited by Spencer Moore. *PLoS One* **12**(5), e0177306 (2017). <https://doi.org/10.1371/journal.pone.0177306>
  50. Lundberg, S.: "Slundberg/Shap." (2020). GitHub Repository. <https://github.com/slundberg/shap>. Accessed: June 2021.
  51. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D., Gebru, T.: Model cards for model reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency-FAT\* '19, pp. 220–229 (2019) <https://doi.org/10.1145/3287560.3287596>.
  52. Moretti, L., Piazza, T.: Transmission of justification and warrant (2013).
  53. Morley, J., Floridi, L., Kinsey, L., Elhalal, A.: From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Sci. Eng. Ethics* (2019). <https://doi.org/10.1007/s11948-019-00165-5>
  54. Mökander, J., Floridi, L.: Ethics-based auditing to develop trustworthy AI. *Mind. Mach.* (2021). <https://doi.org/10.1007/s11023-021-09557-8>
  55. O'Neill, O.: A Question of Trust. Cambridge University Press, Cambridge (2002)
  56. Object Management Group.: Adelard. Macrh 2018. "Structured Assurance Case Metamodel (SACM) Version 2.0."
  57. Owen, R., Bessant, J.R., Heintz, M. (eds.): Responsible Innovation. Wiley, Chichester (2013)
  58. PAIR.: "What-If Tool-People + AI Research (PAIR)." (2020) <https://pair-code.github.io/what-if-tool/>.
  59. Picardi, C., Paterson, C., Hawkins, R., Calinescu, R., Habli, I.: Assurance argument patterns and processes for machine learning in safety-related systems. In: Proceedings of the Workshop on Artificial Intelligence Safety (SafeAI 2020), 23–30. CEUR Workshop Proceedings. CEUR Workshop Proceedings (2020).
  60. Raji, I.D., Smart, A., White, N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., Barnes, P.: Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing 12 (2020)
  61. Rawls, J.: A Theory of Justice, Revised Belknap Press of Harvard University Press, Cambridge (1999)
  62. Reisman, D., Schultz, J., Crawford, K., Whittaker, M.: Algorithmic Impact Assessments: A Practical Framework for Public Accountability. *AI Now* (2018).
  63. Research, IBM.: Introducing AI Fairness 360, A Step Towards Trusted AI. IBM Research Blog (2018). <https://www.ibm.com/blogs/research/2018/09/ai-fairness-360/>.
  64. Royal College of Physicians.: "National Early Warning Score (NEWS) 2." RCP London. (2017). <https://www.rcplondon.ac.uk/projects/outputs/national-early-warning-score-news-2>.
  65. Selbst, A.D., Boyd, D., Friedler, S.A., Venkatasubramanian, S., Vertesi, J.: Fairness and abstraction in sociotechnical systems. In: Proceedings of the Conference on Fairness, Accountability, and Transparency-FAT\* '19. ACM Press, Atlanta, pp. 59–68 (2019) <https://doi.org/10.1145/3287560.3287598>.
  66. Stilgoe, J., Owen, R., Macnaghten, P.: Developing a framework for responsible innovation. *Res. Policy* **42**(9), 1568–1580 (2013). <https://doi.org/10.1016/j.respol.2013.05.008>
  67. Sujan, M., Habli, I.: Safety cases for digital health innovations: can they work? *BMJ Qual Saf*, May, bmjqs-2021-012983 (2021). <https://doi.org/10.1136/bmjqs-2021-012983>.
  68. Sweenor, D., Hillion, S., Rope, D., Kannabiran, D., Hill, T., O'Connell, M.: O'Reilly Media Company Safari. *ML Ops: Operationalizing Data Science* (2020)
  69. Toulmin, S.: The Uses of Argument, Updated Cambridge University Press, Cambridge (2003)
  70. Ward, F.R., Habli, I.: An assurance case pattern for the interpretability of machine learning in safety-critical systems. In: Casimiro, A., Ortmeier, F., Schoitsch, E., Bitsch, F., Ferreira, P. (Eds.) *Computer Safety, Reliability, and Security. SAFECOMP 2020 Workshops*, vol. 12235. Springer International Publishing, Cham, pp. 395–407 (2020). [https://doi.org/10.1007/978-3-030-55583-2\\_30](https://doi.org/10.1007/978-3-030-55583-2_30).

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.