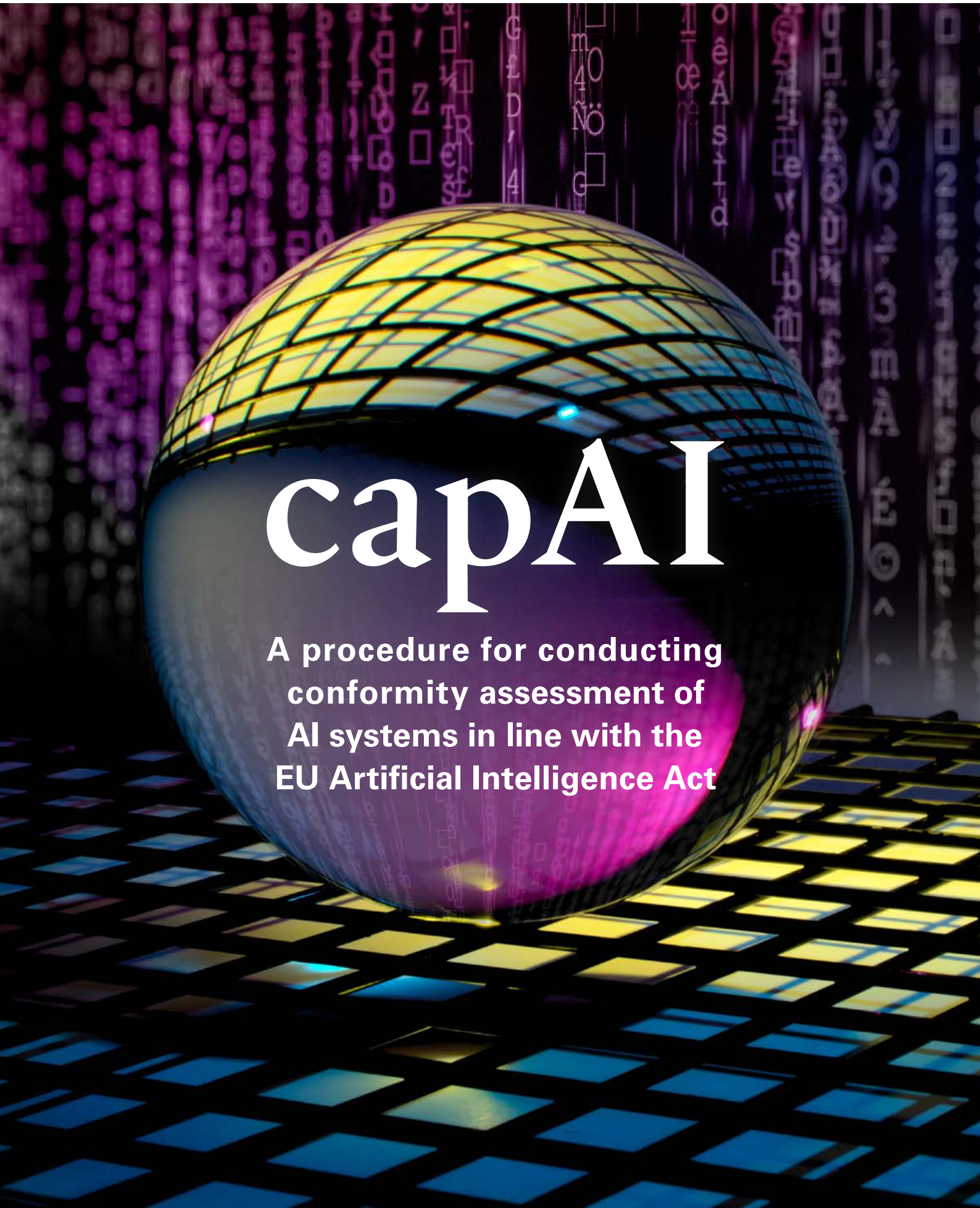

capAI

A procedure for conducting conformity assessment of AI systems in line with the EU Artificial Intelligence Act

Executive summary

Artificial Intelligence (AI) technologies hold great potential for advancing products and services across all industry sectors, society as a whole, and the environment. However, pervasive failures of AI technologies – which can cause harm and fail to meet the normative expectations of citizens and users – undermine trust in such technologies and can hinder their development and use.

To mitigate the risks of AI failures and the lack of trust, careful monitoring of the design, development, and use of AI technologies and assessment of the ethical, legal, and social implications of these technologies are necessary. Indeed, this is the rationale underpinning the EU's Artificial Intelligence Act (AIA), which has been drafted as a set of harmonised rules to aid organizations in designing trustworthy AI systems. Central herein is the conformity assessment of high-risk AI systems. Proactively assessing AI systems can prevent harm by avoiding, for example, privacy violations, discrimination, and liability issues, and in turn, prevent reputational and financial harm from organisations that operate AI systems.

We have developed capAI, a conformity assessment procedure for AI systems, to provide an independent, comparable, quantifiable, and accountable assessment of AI systems that conforms with the proposed AIA regulation. By building on the AIA, capAI provides organisations with practical guidance on how high-level ethics principles can be translated into verifiable criteria that help shape the design, development, deployment and use of ethical AI. The main purpose of capAI is to serve as a governance tool that ensures and demonstrates that the development and operation of an AI system are trustworthy – i.e., legally compliant, ethically sound, and technically robust – and thus conform to the AIA.

capAI provides a structured process for ensuring and demonstrating adherence to defined organizational values. To achieve this goal, capAI adopts a process view of AI systems by defining and reviewing current practices across the five stages of the AI life cycle: design, development, evaluation, operation, and retirement. capAI enables technology providers and users to develop ethical assessment at each stage of the AI life cycle and to check adherence to the core requirements for AI systems set out in the AIA. The procedure consists of three components:

1. an internal review protocol (IRP), which provides organisations with a tool for quality assurance and risk management;
2. a summary datasheet (SDS) to be submitted to the EU's future public database on high-risk AI systems in operation; and
3. an external scorecard (ESC), which can (optional) be made available to customers and other stakeholders of the AI system.

By following the IRP, organisations can conduct conformity assessment in line with, and create the technical documentation required by, the AIA. It follows the development stages of the AI system's lifecycle, and assesses the organisation's awareness, performance, and resources in place to prevent, respond to and rectify potential failures. The IRP is designed to act as a document with restricted access. However, like accounting data, it may be disclosed in a legal context to support business-to-business contractual arrangements, or as evidence when responding to legal challenges related to the AI system audited.

The SDS is a high-level summary of the AI system's purpose, functionality and performance that fulfils the public registration requirements, as stated in the AIA.

The ESC is generated through the IRP and summarises relevant information about the AI system along four key dimensions: (1) purpose, (2) values, (3) data, and (4) governance. It is a public reference document that should be made available to all counterparties concerned.

Conjointly, the internal review protocol and external scorecard provide a comprehensive audit that allows organisations to demonstrate the conformance of the AI system with the EU's Artificial Intelligence Act to all stakeholders.

The **capAI** procedure will be updated regularly to reflect changes in the proposed regulation, as well as to incorporate potential improvements. The current version will be available for download at SSRN.

The version number and release date for the underlying report are:

Version 1.0
March 23, 2022

Table of Contents

1 Conformity assessment for trustworthy AI 1

1.1 Delivering on the promise that AI holds 1
1.2 When does the AIA conformity assessment mandate apply? 2
1.3 The key components of trustworthy AI systems 4
1.4 Using ethics-based auditing to operationalise conformity assessment 5
1.5 How to use this document 7

PART I – THE PROTOCOL 8

2 The requirements under the AIA in brief 8

2.1 What is the objective of the AIA? 8
2.2 What systems are covered? 9
2.3 Who needs to act? 9
2.4 What actions are needed to comply? 10
2.5 What are the penalties for non-conformance? 13

3 The conformity assessment procedure 14

3.1 When to use this procedure 14
3.2 How to use capAI 16
3.3 The outputs of the capAI process 16
3.4 Key actors 17
3.5 High-level navigation 17

4 Internal review protocol 18

4.1 Stage 1: Design 18
4.2 Stage 2: Development 20
4.3 Stage 3: Evaluation 22
4.4 Stage 4: Operation 24
4.5 Stage 5: Retirement 26

5 Summary datasheet 27

6 External scorecard 28

6.1 Content of the external scorecard 28
6.2 Graphical representation of the external scorecard 29

PART II – THE REFERENCE 30
7 Defining the AI process flow 30
7.1 Defining AI development and operation as a process 30
7.2 The five key stages in the AI life cycle 33
7.3 The design stage 33
7.4 The development stage 39
7.5 The evaluation stage 49
7.6 The operation stage 53
7.7 The retirement stage 55
8 The rationale for ethics-based auditing of AI systems 56
8.1 Prevalence and modes of AI ethics failure 56
8.2 Conformity assessment and post-market monitoring as stipulated in the AIA 58
8.3 Roles and responsibilities in an emerging European auditing ecosystem 61
8.4 The remaining gap 63
9 The implementation of ethics-based auditing 64
9.1 Introduction 64
9.2 Defining key terms 64
9.3 Background 65
9.4 How capAI harnesses the promise of ethics-based auditing 68
9.5 Best practices for successful implementation 69
9.6 Managing known limitations and pitfalls 71
10 Concluding remarks 74
Glossary of key terms 75
References 77

1 Conformity assessment for trustworthy AI

1.1 Delivering on the promise that AI holds

Artificial intelligence (AI) holds great promise to support human flourishing, economic prosperity and sustainable growth. Enabled by advances in machine learning, access to computing power at decreasing costs, the growing availability of data, and the ubiquity of digital devices, AI is set to benefit the public, private and third sectors alike. AI has become a growing resource of interactive, autonomous and self-learning agency, which can perform tasks that would otherwise require human intelligence and intervention to be successfully executed [1]. Delegating tasks to AI systems can help increase consistency, improve efficiency and increase access to a service or product. Recent estimates suggest that AI may boost global GDP by around 15% by 2030 [2]. The positive impact of AI is not just economic but also social [3]. Consider, for example, the application of AI in the healthcare sector, where AI-powered image recognition enhances diagnostic services, or in the public sector, where AI is used to improve the quality of community services through more accurate forecasting [4]. Due to its ability to draw inferences from large and even less-structured data, AI is a particularly useful tool for enabling new solutions to complex problems, such as achieving the *UN Sustainable Development Goals* (SDGs) [5].

However, delegating tasks to AI systems is also coupled with ethical challenges [6]. AI systems may enable malfeasance, reduce human control, remove human responsibility, devalue human skills, and erode human self-determination [7]. This is why, for public and private actors seeking to reap the benefits of AI, it is essential to understand and address the ethical implications of AI. This call has been widely heeded, and numerous documents have been published by public and private actors that stipulate principles for how to design and deploy ethical AI, first and foremost the EU's *Ethics Guidelines for Trustworthy AI* [8], but also the OECD's *Recommendation of the Council on Artificial Intelligence* [9]. Although varied in their terminology, different high-level guidelines tend to converge around five principles: beneficence, non-maleficence, autonomy, justice and explicability [1]. These guidelines were embedded in, and form the foundation of, the proposal for the EU's *Artificial Intelligence Act* [10]. The convergence into one piece of regulation is promising with respect to the value of the adopted principles, and how to manage the normative tensions between high-level principles. Consider, for example, the uncertainty in prioritising between conflicting definitions like 'equality of opportunity' and 'equality of outcome'. This indeterminacy hinders the translation of AI ethics principles into practices and leaves room for unethical behaviours like 'ethics shopping', i.e., mixing and matching ethical principles from different sources to justify some pre-existing behaviour; 'ethics bluewashing', i.e., making unsubstantiated claims about AI systems to make them appear more ethical than they are; and 'ethics lobbying', i.e., exploiting ethics to delay or avoid good and necessary legislation [11]. By building on the AIA, **capAI** provides organisations with

practical guidance on how high-level ethics principles can be translated into verifiable criteria that help shape the design, development, deployment and use of ethical AI.

capAI defines a procedure to implement e*thics-based auditing* [12, 13], offers the most effective approach to conduct a conformity assessment in line with the AIA, as it identifies and enables the correction of unethical behaviours of AI systems, and informs ethical deliberation throughout the process of designing such systems [14]. A wide-scale application of such ethics-based audits in practice may prevent ethics-based AI failures, improve chances of rectifications, and provide a transparent and reasonable ground for an explanation of such failures if they occur. **capAI** can guide any organisation to design better AI systems by raising the key ethics questions and by requiring explicit statements where trade-off decisions must be made across the AI life cycle. The premise of **capAI** is as simple as it is powerful: by adopting an auditable and standardised process for developing, deploying and operating AI, the most common failure modes can be avoided.

In the remainder of this chapter, we outline our approach to ethics-based auditing, and how to use **capAI** in practice. The subsequent *protocol section* provides the details about how to use the IRP and how it addresses the reporting requirements under the AIA. Finally, the *reference section* provides further explanations and details, for those who wish to dive deeper into the different features of **capAI**. A glossary of key terms and the references used in creating the protocol are provided at the end of this document.

1.2 When does the AIA conformity assessment mandate apply?

The main objective of the AIA is to ensure that AI systems within the EU are safe, and comply with existing law on fundamental rights, norms and values. The AIA defines AI systems broadly by including logic- or rule-based information processing (such as expert systems), as well as probabilistic algorithms (such as machine learning). Like the GDPR, it applies to all firms wishing to operate AI systems within the EU, irrespective of whether they are based in the EU or not. The AIA adopts a *risk-based approach* to regulating AI systems. This is an important aspect, as the regulatory requirements differ, based on the level of risk foreseen for a given AI system. In terms of their perceived risk, some AI systems are banned outright, while others are not regulated at all. The following provides a brief overview of the main risk categories; we will delve into the AIA in Section 2.1 of this document.

First, there are ‘prohibited AI practices’, which are banned outright. These include real-time biometric systems (with a few exceptions for law enforcement purposes, for example) and social scoring algorithms. It also makes special provisions for AI systems that may involve manipulation risks, such as chatbots or deepfakes.

Second, there are ‘high-risk AI systems’, i.e., AI systems employed in areas specifically listed as high-risk in the AIA, such as law enforcement, management of

critical infrastructure or recruitment.[1] For these systems, there is a complex compliance regime regarding their development and operation. It is here that **capAI** helps organisations comply in their development and use.

Third, there are 'low-risk AI systems', because they neither use personal data nor make any predictions that influence human beings. According to the European Commission, most AI systems will fall into this category. A typical example is industrial applications in process control or predictive maintenance. Here, there is little to no perceived risk, and as such, no formal requirements are stipulated by the AIA.

The following flowchart summarises the conformity assessment requirements:

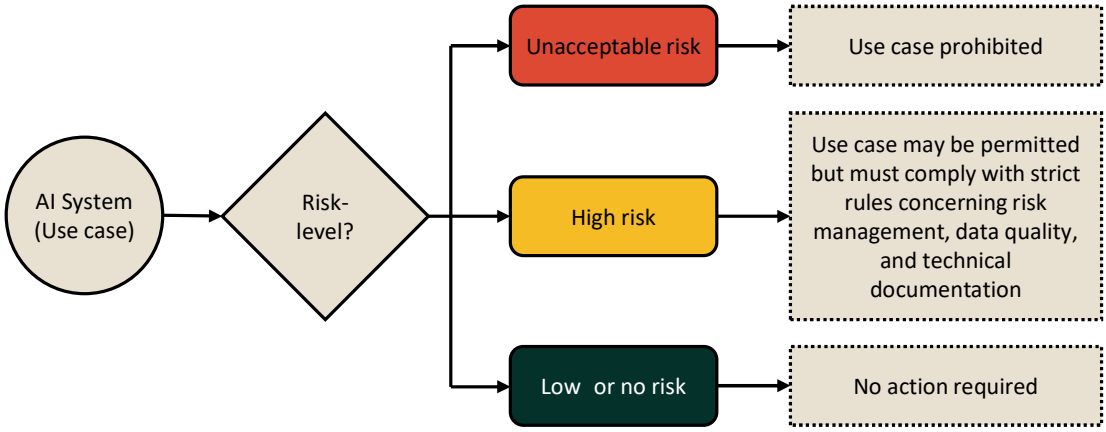

Figure 1: Risk categories for AI use cases under the AIA [14]

It is important to note that the requirements stipulated in the AIA apply to all high-risk AI systems. However, the need to conduct conformity assessments only applies to 'standalone' AI systems. For algorithms embedded in products where sector regulations apply, such as medical devices, the requirements stipulated in the AIA will simply be incorporated into existing sectoral testing and certification procedures. Nonetheless , it is highly recommended to use **capAI** to assess the AI component of these products to ensure a comprehensive product safety assessment.

Finally, it is also a good practice to use the protocol even for low-risk AI applications that are currently not covered by the AIA. After all, there is always room for more post-compliance, ethical behaviour.

[1] The full list of high-risk AI systems is found in ANNEX III to the AIA.

1.3 The key components of trustworthy AI systems

The AIA builds on the work of the EU High-Level Expert Group (HLEG)[2] that has set out the principles for **trustworthy AI**, defined through its three components. AI systems should be:

1. **Lawful**,
2. **Ethical**, and
3. **Technically robust**.

Confusingly, however, many other terms are also being used in the EU AIA, such as 'safe' or 'transparent'. The reason for the diversity in terminology is that these are interrelated. For example, a lack of technical robustness can lead to ethical concerns like bias, which in turn can lead to legal consequences in terms of discrimination. The main legal risks stem from privacy violations (under GDPR rules) and bias, both of which also carry significant reputational and ethical risks. Likewise, the lack of explainability also has legal, ethical and reputational risks [15]. Thus, the legal, ethical and technical aspects of AI systems' performance are closely intertwined. A central issue, at the core of this definitional problem, is that ethical principles stand in complex relationships that often require the management of trade-offs [16]. Consider, for example, the use of an AI system to determine the car insurance premium for a customer: while women are known to be safer drivers, it is not possible to use gender as a variable as this would discriminate against male applicants. Thus, the criteria of fairness on the one hand, and inclusivity on the other, need to be carefully balanced.

Given this complexity of relations, **capAI** adopts a post-compliance, ethics-centred approach, as the adherence to ethical norms and values is considered to provide the highest standards, covering main problems such as privacy, bias and explainability. In and of itself, the proposed EU legislation constitutes *hard* governance. However, the AIA also leaves room for *soft* governance in general and post-compliance, ethics-based auditing in particular [17].[3] These hard and soft mechanisms often complement and mutually reinforce each other [18]. In the case of AI governance, this is especially true since laws may not always be up to speed in sectors that experience fast-paced technological innovation. Further, decisions made by AI systems may deserve scrutiny even when they are not illegal. Hence, there is always room for post-compliance, ethics-based governance whereby organisations can prove adherence to voluntary standards that go over and above existing regulations.

---

[2] Disclosure: Professor Luciano Floridi was a member of the EU HLEG. Also, Professor Mariarosaria is the chair of the board of directors of Noovle S.p.A.

[3] *Hard governance* refers to systems of rules elaborated and enforced through institutions to govern agents' behaviour. In contrast, *soft governance* embodies mechanisms that exhibit some degree of contextual flexibility, like subsidies and taxes.

This approach is illustrated in the 'sand cone' chart, which visualises the approach of 'cumulative capabilities' promoted by Ferdows and De Meyer in the context of manufacturing [19]. The argument is that capabilities are not independent but, rather, build on one another. In this context, we refer to legal compliance as the foundation upon which trustworthy AI systems can be developed. AI systems need to demonstrate technical robustness, that is, to demonstrate that they perform to expectations under all conceivable scenarios. Once these two conditions are met, one can assess the adherence to ethical standards. This third and last requirement is conceptually the hardest, as it is a largely qualitative evaluation of how ethical trade-offs are managed, and of the degree to which these trade-offs adhere to ethical norms and values.

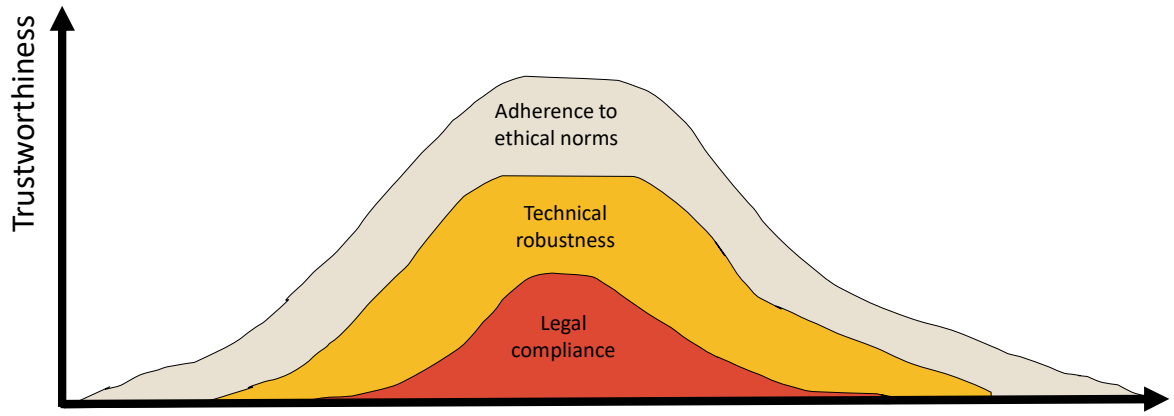

Figure 2: The sand cone model of cumulative capabilities, applied to AI trustworthiness

Source: adapted from [19].

1.4 Using ethics-based auditing to operationalise conformity assessment

The concept of auditing is widely established in financial accounting, software development and beyond. In this context, we refer to auditing as a structured process through which organisations and AI systems are assessed for consistency with relevant principles or norms. Further, we use the expression 'ethics-based' instead of 'ethical' to avoid any confusion. We do not refer to a kind of auditing done ethically, or to the ethical use of AI in auditing, but to an auditing process that assesses AI systems based on their adherence to predefined ethics principles. So defined, ethics-based auditing shifts the focus of the discussion around AI ethics from the abstract to the operational. It is also compatible with current best practices in agile software development, insofar as steps are taken to ensure ethical alignment throughout the product life cycle, thereby permeating the conceptualisation, design, deployment and use of AI. The logic of process thinking applied here is already widely established in the industry, where it underpins both quality management and productivity gains in many sectors, from the automotive industry to healthcare [20].

capAI has been developed with two use cases in mind. First, providers of 'high-risk' AI systems may use **capAI** to demonstrate compliance with the EU's AIA. Second, providers of 'low-risk' AI systems, i.e., systems that do not fall within the regulatory scope of the AIA, may use **capAI** to operationalise their commitments to voluntary codes of conduct. Let us consider these two use cases in turn.

The AIA requires providers of high-risk AI systems to conduct conformity assessments before placing their product or service on the European market. Occasionally, such conformity assessments may need to be performed with the involvement of an independent third-party body. Yet, for most AI systems, conformity assessments based on 'internal control' will be sufficient. However, while the AIA stipulates a wide range of procedural requirements for conformity assessments based on internal control, it does not provide any detailed guidance on how these requirements should be implemented in practice. **capAI** satisfies all requirements for conformity assessments based on internal control. It thereby constitutes a ready-to-go and easy-to-use procedure for providers of high-risk AI systems to demonstrate compliance with the AIA.

Even if not all AI systems are covered by the proposed European legislation (see Section 1.2), all technology providers have an interest in ensuring that the AI systems they design and deploy are not only legal but also ethical and technically robust. Therefore, many organisations have drafted and committed themselves to different sets of 'ethical principles' to guide the design and use of AI systems. Unfortunately, practitioners have so far lacked both incentives and tools to translate abstract principles into best practices to ensure and validate that AI systems are ethically sound. **capAI** addresses this gap. By following our protocol, organisations can validate claims about the AI systems they design and deploy, thereby operationalising their commitments to voluntary codes of conducts.

The fundamental idea behind **capAI** is that ethics-based auditing should help stakeholders identify and communicate the normative values embedded in AI systems. It thereby informs ethical deliberation among AI practitioners, and enables public discourse on what makes socially acceptable use of AI systems. If successfully implemented following the best practices presented in this protocol, ethics-based auditing can help organisations manage the ethical risks posed by AI while allowing society at large to reap the full economic and social benefits of automation.

Ethics-based auditing is a governance mechanism that can be used by organisations that design and deploy AI systems to control or influence the behaviour of AI systems. Operationally, ethics-based auditing is characterised by a structured process through which an entity's current or past behaviour is assessed for consistency with relevant principles or norms. Note that, while AI should also be lawful and technically robust, our focus here is post legal compliance, and is centred on the ethical aspects (see Section 1.3).

As a governance mechanism, auditing has a long history of promoting trust and transparency in areas like security and financial accounting. Based on experiences

from these domains, two transferable lessons can be drawn. First, the process of auditing is always purpose-oriented. This means that ethics-based auditing differs from merely publishing a code of conduct in that it aims to demonstrate adherence to a predefined baseline. Second, auditing presupposes operational independence between the auditor and the auditee. Whether the auditor is a government body, a third-party contractor, an industry association or a specially designated function within larger organisations, the main point is to ensure that the auditing runs independently of the day-to-day management of the audited organisation. Note that using a 'notified body', i.e., a third-party auditor, does not apply to all high-risk AI system cases. A detailed description of different stakeholders' roles and responsibilities involved in ethics-based auditing is given in Section 5.

1.5 How to use this document

This document comprises two main parts. In Part I – *The Protocol* – Chapter 2 summarises the main aspects of relevance within the AIA. Chapter 3 then introduces the internal review protocol (IRP) and provides a corresponding checklist for each stage of the AI life cycle. A subset of the information collated for the internal review protocol then comprises the summary datasheet (SDS) and optional external scorecard (ESC), which are discussed subsequently. Part II – *The Reference* – discusses the AI workflow in detail. Chapter 4 provides a theoretical foundation for our work. Readers familiar with machine learning techniques and workflow may wish to skip it, and use it as a reference only. Chapter 5 provides the detailed justification for using an ethics-based approach to assessing the conformity of AI systems. Finally, Chapter 6 outlines in detail how **capAI** can help organisations to ensure and demonstrate adherence to the requirements on AI systems stipulated in the AIA.

PART I – THE PROTOCOL

2 The requirements under the AIA in brief

In this section, we summarise the role of conformity assessment within the context of the AIA. In doing so, we highlight the set objectives of the AIA and clarify which systems are covered by the AIA, who needs to act upon it, what actions are required, and what potential penalties await non-conformance. This section constitutes a necessary simplification of a complex legal document; readers are strongly advised to consult the actual legal texts of the AIA and its Annexes.

2.1 What is the objective of the AIA?

The proposed AI regulation seeks to ensure that any AI system operated within the EU, or affecting EU citizens, is trustworthy – defined as legally compliant, technically robust and ethically sound. Essentially, the AIA seeks to make sure that AI systems comply with existing EU laws, rights and values to prevent harm to its citizens. To do so, the AIA adopts a risk-based approach, which is essential to understand, because much of the assessment and documentation requirements set out for operators (organisations that develop or commission AI systems) are linked to the respective risks these systems may entail. Much of the assessment is based on risk identification, mitigation and eradication. The following table provides an overview of those aspects of the AIA and its Annexes that are directly relevant to conducting conformity assessments.

Title I	Scope and definitions
Title II	Prohibited AI practices
Title III	High-risk AI systems
Title IV	Transparency obligations for certain AI systems
Title V	Measures in support of innovation
Title VI	Governance
Title IX	Codes of conduct (for low-risk AI systems)
Annex I	Artificial intelligence techniques and approaches
Annex IV	Technical documentation
Annex V	EU declaration of conformity
Annex VI	Conformity assessment procedure based on internal control
Annex VII	Conformity based on assessment of quality management system and assessment of technical documentation

Table 1: *An overview of the relevant sections of the AIA*

2.2 What systems are covered?

The AIA takes a very inclusive view of what are considered to be AI systems, which are defined as software that is developed using a set of techniques (set out in Annex I), which include machine learning and other statistical approaches (probabilistic systems), and expert systems (rule-based or deterministic systems). The definition is much broader than conventional definitions of AI, which are often restricted to machine learning using supervised, unsupervised and reinforcement learning methods. Also covered under the AIA are other predictive analytics, such as Bayesian estimators, as well as deterministic expert systems that have been in operation for many decades in a wide range of contexts. In summary, the AIA covers the following systems and approaches:

1. **Machine-learning approaches**, including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning;

2. **Logic- or knowledge-based approaches**, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems;

3. **Statistical approaches**, Bayesian estimation, search and optimisation methods.

The rationale for adopting such a broad definition is that the actual system is less of a concern than the use case. In other words, the AIA takes a **broad view of what constitutes an AI system**, but a **narrow view concerning its use case**. This view is consistent with the risk-based approach that the AIA adopts. It is largely agnostic to the means and focuses on the risk inherent in any prediction that can potentially impact a person's wellbeing, more or less directly. Consequently, many existing expert systems that have been in operation for many years, and even decades in some cases, will now be under scrutiny. However, the AIA does make provision for systems that are conforming to 'harmonised standards', allowing operators to rely on a 'presumption of conformity'. The latter is likely to apply to AI systems already covered by other EU legislation, such as sector-specific regulations for medical devices or toys. In any case, the onus is on the provider of any AI system to investigate to what degree they must act.

2.3 Who needs to act?

The AIA identifies a wide range of roles that organisations can take within its context. These include, for example, providers, authorised representatives, importers, distributors and users. Not all of these roles come with the same obligations under the AIA. It is also important to state that the AIA has extra-territorial reach. It applies to

any firm operating an AI system within the EU and firms located outside the EU. To ensure the AI systems used within the EU market conform to the AIA, the main onus rests with (a) the providers, who place an AI system on the EU market or put into service an AI system for use in the EU market; (b) users located within the EU market; and (c) providers or users of AI systems that are located outside of the EU, but whose system is used (or has an output) on the EU market.

capAI is designed specifically for providers seeking to conduct conformity assessments in line with the requirements stipulated in the AIA. Still, where appropriate, references are made to how other actors (such as users, independent third parties or providers of non-high-risk AI systems) can use the different components of **capAI**. Irrespective of the legal mandate, we argue that following the IRP and/or publishing an ESC, denote good practice for low-risk AI systems even where such procedures are not mandated. Software vendors (who are not providers themselves) in particular may choose to do so. Similarly, even where the AI system in question is embedded in a wider system that is covered under a sectoral regulation, **capAI** can help providers demonstrate conformity with the broader trustworthiness mandate.

2.4 What actions are needed to comply?

If an AI system falls under the AIA, as outlined above, then the actions needed are determined by the level of risk embedded in the respective system. Thus, the initial question for providers is to determine that risk level, in light of the types and categories set out in the AIA:

- **Prohibited practices** denote the highest risk category, and these systems are banned outright. These include:
    - **Real-time biometric systems** that can be used for any type of surveillance, although exceptions do apply here for crime prevention and criminal investigations in law enforcement and national security contexts.
    - **Social scoring algorithms** that can be used to evaluate individuals based on personal characteristics and/or their behaviour in a manner that could cause harm or lead to unfavourable treatment of that individual.
    - **Manipulative systems** that exploit the vulnerabilities of specific individuals to distort their behaviour in a manner that is likely to cause physical or psychological harm.
- **High-risk AI systems**, listed in Annex III and likely to constitute the majority of AI systems. These include:

- **Biometric identification and categorisation of natural persons**, to the extent these do not fall under the aforementioned prohibited practices.

- **Management and operation of critical infrastructures**, such as AI systems used in safety-relevant components of the management of utilities and traffic.

- **Education and vocational training,** such as AI systems used to assess students in educational settings, or assign people to training offerings.

- **Employment and worker management,** such as AI systems used for the recruitment or assessment of employees, including questions such as promotion, performance management and termination.

- **Access to essential services**, such as AI systems that govern the access to private and public sector services and related actions, including the assessment of creditworthiness, credit scoring, or establishing the order of priority of access to such services. (Note: this aspect applies particularly to AI systems used in the financial services sector).

- **Law enforcement**, which includes a broad range of AI systems used, among other things, to assess the risk of any individual committing an offence, or of re-offending; predicting the likelihood of criminal offences (e.g., predictive policing and profiling), as well as the detection and investigation of fraudulent content;

- **Border control management**, including AI systems used for the control and management of borders, migration and asylum processes, such as validating travel documents and assessing the eligibility for asylum.

- **Administration of justice and democratic processes**, including any AI system used to assist in the judicial process by assessing and interpreting facts, and/or making legal recommendations in response to facts.

- **Low-risk AI systems**, which include systems that neither use personal data nor make predictions that are likely to affect any individual directly or indirectly, like industrial applications in predictive maintenance.

- **Embedded AI systems**, which are components of products or services covered under other EU regulations, such as for toys or medical devices. While these systems do not fall under the AIA, they will still have to be compliant with the requirements set out in the AIA under the harmonisation directive**.**

For AI systems that are not prohibited, but are low-risk and not covered under existing sectoral regulation, the rules for ‘high-risk AI systems’ will apply. These

systems have to undergo conformity assessments**.** However, the conformity assessment can be conducted in different ways. For example, some AI systems are part of consumer products that are already subject to testing and certification before they can be placed on the market (such as medical devices). For these embedded AI systems, no extra conformity assessment procedure will be necessary. Instead, the requirements stipulated in the AIA will be incorporated into existing sector-specific safety legislation.

In contrast, 'standalone' high-risk AI systems have to undergo an AI-specific conformity assessment before they can be placed on the EU market. There are two ways to conduct such conformity assessments: **conformity assessment based on internal controls** (see Annex VI), and in some cases, a **conformity assessment of the quality management system and technical documentation conducted by a third party**, referred to as a 'notified body' (see Annex VII). These are two fundamentally different conformity assessment procedures. The type of procedure required for a specific AI system is outlined in Annex III, and depends on the use case, i.e., purpose, for which it is employed. In short, high-risk AI systems that use biometric identification and categorisation of national persons (Annex III, point 1) must conduct a third-party conformity assessment.[4] For most high-risk AI systems, however, conformity assessment using internal controls will be sufficient. **capAI** has been specifically developed to help providers of standalone high-risk AI systems to conduct conformity assessments based on internal controls in line with the requirements set out in the AIA.

Providers of AI systems that interact directly with humans – chatbots, emotional recognition, biometric categorisation and content-generating ('deepfake') systems – are subject to further transparency obligations. In these cases, Title IV, Article 52, in the AIA requires providers to make it clear to the users that they are interacting with an AI system and/or are being provided with artificially generated content. The purpose of this additional requirement is to allow users to make an informed choice as to whether or not to interact with an AI system and the content it may generate.

In summary, the AIA sets out a complex set of requirements for conformity assessment for high-risk AI systems in relation to (1) which use case category under Titles III and IV it falls under, and (2) the degree to which the AI system in question may be covered by existing legislation already. We assess that, in most cases that do not use any biometric data, a conformity assessment using internal controls will be required. However, organisations are advised to assess carefully which procedure is required in their respective case.

[4] Exemptions apply where harmonised standards have been applied in full, in which case the provider can choose to either implement the conformity assessment using internal controls, or use external controls via a third-party notified body.

2.5 What are the penalties for non-conformance?

The penalties set out in the AIA for non-conformance are, in principle, very similar to those set out in the GDPR. The main thrust is for penalties to be effective, proportionate and dissuasive. The sanctions are structured in a similar way to those under the GDPR, and include three main levels:

- Non-compliance with regard to prohibited AI practices, and/or the data and data governance obligations set out for high-risk AI systems can incur a penalty of up to €30m, or 6% of total worldwide turnover in the preceding financial year (whichever is higher).

- Non-compliance of an AI system with any other requirement under the AIA than stated above can incur a penalty of up to €20m, or 4% of total worldwide turnover in the preceding financial year (whichever is higher).

- Supply of incomplete, incorrect or false information to notified bodies and national authorities in response to a request can incur a penalty of up to €10m, or 2% of total worldwide turnover in the preceding financial year (whichever is higher).

It should be noted that the enforcement of the AIA sits with the competent national authorities. Individuals adversely affected by an AI system may have direct rights of action, for example, concerning privacy violations or discrimination.

3 The conformity assessment procedure

3.1 When to use this procedure

The AIA sets out extensive requirements for AI systems according to the level of risk these pose to the wellbeing of EU individuals. The conformity assessment proposed in the AIA is the key enforcement mechanism to ensure that providers adhere to these requirements. However, while the AIA provides extensive discussion of the aspects and outcomes of AI systems that it seeks to prevent, it **neither prescribes nor details the form of such conformity assessments**. This is the gap that **capAI** aims to fill. Specifically, **capAI** seeks to aid firms required to conduct an internal conformity assessment of high-risk AI systems (Sections 4 and 5 in the reference section detail the justification and implementation of our approach). Beyond this mandate, however, different components of **capAI**, such as the IRP and the ESC, can also be used by providers of low risk and embedded AI systems.

High-risk AI systems – internal control (Title III and Annex VI)

capAI is explicitly designed to help organisations implement conformity assessment using an internal control (Annex VI), which in our view is the most common case. However, the requirements under the AIA are extensive. **capAI** has been designed to help organisations to fulfil only the following requirements (marked green in Figure 3):

✓	The conformity assessment of high-risk AI systems (Article 43)
✓	The technical documentation of the AI system, detailing its objectives and functionality
✓	A summary datasheet for submission to the planned EU national database
✗	A system for post-launch monitoring and logging of key events
+	Optional: an external scorecard to be made publicly available to customers of, and counterparties to, the AI system in question

Figure 3: *capAI coverage for high-risk AI systems, internal control*

High-risk AI systems – external control (Title III and Annex VII)

As stipulated in Annex VII to the AIA, specific high-risk AI systems will need to be assessed by a notified body (external auditor). **capAI** can be used by the notified body in the same way as denoted above for conformity assessment with internal control, and will provide the following (marked green in Figure 4):

✓	The conformity assessment of high-risk AI systems (Article 43)
✓	The technical documentation of the AI system, detailing its objectives and functionality
✓	A summary datasheet for submission to the planned EU national database
✗	A quality management system for the AI system in question
✗	A system for post-launch monitoring and logging of key events
+	Optional: an external scorecard to be made publicly available to customers of, and counterparties to, the AI system in question

Figure 4: *capAI coverage for high-risk AI systems, external control*

Low-risk and embedded AI systems (Title IX)

Low-risk AI systems and those embedded in products or services regulated under further sectoral regulation do not fall under the conformity assessment requirements stipulated by the AIA. However, under the harmonisation mandate, embedded AI systems will be held to the same standard as standalone AI systems covered by the AIA. Thus, we strongly recommend using **capAI** as best practice in the form of a *voluntary code of conduct* for low-risk and embedded AI systems.

Providers of low-risk AI systems should draw up and apply voluntary codes of conduct related to their internal procedures and the technical characteristics of the systems they design and deploy. The critical difference between these voluntary codes of conduct and the other requirements stipulated in the AIA is that the former focus on *process management* rather than *goal management*. This leaves individual organisations free to draw up guiding principles of their own, or adopt guidelines recommended by the European Artificial Intelligence Board, or declare adherence to any other set of standards relevant for their specific industry or use case. The main takeaway here is that providers of low-risk AI systems, i.e., systems that do not fall within the regulatory scope of the AIA, can use **capAI** to operationalise their commitments to voluntary codes of conduct. To do so, providers of these systems can implement the same procedures described in Section 2 above to improve their internal quality management systems. The only difference is that they are not legally obliged to sign an EU declaration of conformity. Figure 5 highlights how **capAI** can support providers of low-risk AI systems.

➕	Optional: Adherence to a voluntary code of conduct
➕	Optional: Technical documentation of the AI system, detailing its objectives and functionality
➕	Optional: an external scorecard to be made publicly available to customers of, and counterparties to, the AI system in question
❌	A system for post-launch monitoring and logging of key events

Figure 5: capAI coverage for low-risk AI system

3.2 How to use capAI

capAI implements the principles of ethics-based auditing (detailed in Sections 4 and 5) and applies them to each of the five key stages of the AI life cycle. At each stage, the crucial ethical issues are addressed, bringing to light the key tensions that underpin the main ethical issues. It is acknowledged that these are trade-offs. In other words, there is no optimal solution that will comprehensively address all ethical concerns. The procedure adopts a process view of AI systems by defining and reviewing current practices across the concept, development, evaluation, operation and retirement stages of the AI life cycle. The approach sets out a continuous, holistic, dialectic and traceable process that – at each stage – identifies the core requirements for AI systems to adhere to the ethical principles set out in the EU HLEG guidelines.

3.3 The outputs of the capAI process

capAI provides three documents that organisations can use when in the process of ensuring and demonstrating adherence to the AIA: (1) an internal review protocol (IRP), which provides organisations with a management tool for quality assurance and risk management; (2) the summary datasheet (SDS) to be submitted to the EU database; and (3) an optional external scorecard (ESC), which should be made available to customers and other stakeholders of the AI system.

The IRP follows the development stages of the AI system's life cycle, and helps organisations to assess the awareness, performance and resources in place to prevent potential failures, as well as the process for responding and rectifying potential failures. The IRP is designed to act as a document with restricted access, yet, like accounting data, may be disclosed in a legal context to support business-to-business contractual arrangements, or as evidence when responding to legal challenges related to the AI system being audited. The SDS synthesises key information about the AI system, including its purpose, status and key contact details to providers and relevant representatives. The ESC is generated through the IRP. It summarises relevant information about the AI system into an overall risk score. System details are provided

for four key dimensions: (1) purpose and ethics norms, (2) data and privacy, (3) bias and explanation, and (4) governance and rectification. It is a public reference document that organisations can make available to different stakeholders on a voluntary basis. Conjointly, the IRP and ESC provide comprehensive documentation that allows organisations to show to all stakeholders the conformity of a specific AI system to the requirements stipulated in the AIA.

3.4 Key actors

The procedure for generating both the IRP and the SDS is outlined below by defining a set of questions for each of the key actors involved at different stages in the AI life cycle. These stakeholders are:

1. **Top manager responsible for AI**, who bears responsibility for justifying the application and performance of the AI system to all stakeholders, internally and externally.
2. **Product owner**, who is responsible for the performance of the AI system in question.
3. **Project manager**, who leads the development (or, if externally sourced, procurement) process.
4. **Data scientist**, who leads the technical implementation of the AI system in question.

In what follows, we cover the IRP and ESC in sequence. As the ESC requires auditing summary data of the IRP, the IRP needs to be complete before assembling the ESC.

3.5 High-level navigation

The IRP follows the AI process flow, which is outlined in detail in Section 4, and addresses the main ethical issues and resulting technical considerations that arise at each stage, as illustrated in Figure 6:

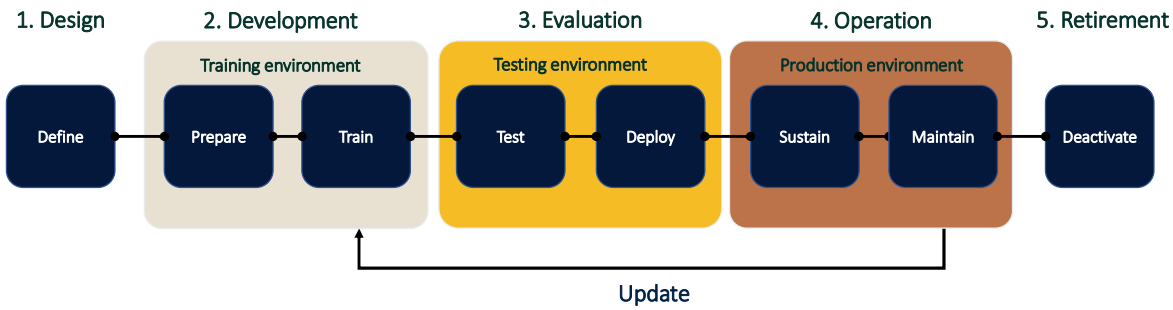

Figure 6: AI process flow with its five stages and key steps

At each stage, the requirements consist of two aspects: (1) organisational governance, and (2) the use case for the AI system in question. Each requirement is linked to an actor, who is best placed to ensure and confirm that the requirement in question is met. For many requirements, supporting evidence will be requested. Overall, there are 40 items to complete in the protocol.

The IRP follows the five stages of the AI process flow outlined above. It is suggested the IRP be treated as a 'check-list' and completed chronologically – from the design stages to the retirement stage. Please note that the reporting requirements vary significantly across the stages, as the most critical decisions (that will determine the actual performance of the system) are made during the earlier stages of the AI life cycle.

4 Internal review protocol

4.1 Stage 1: Design

AI systems may outperform humans in decision-making across many domains, yet, it remains superfluous – even unsuitable – for many purposes [21, 22]. To distinguish when an AI system is fit for purpose, organisations should start any AI development with a Concept stage, that is, a stage for eliciting the use case's requirements (technical and ethical) and users' expectations of a product. The Concept stage serves two goals. First, it prevents project misspecification, that is, a situation where the AI system is unreflective of the underlying problem [23]. Second, it facilitates a feasibility assessment, which is a study of the system viability, limitations and trade-offs [24]. Failure to meet any of these goals will result in an AI that malfunctions [25] or unintentionally reinforces existing societal disparities [23].

Addressing the need to embed ethical considerations into the design of AI [26, 27], we propose that the concept stage must include a definition of both organisational governance [28, 29], and the use case's functional requirements. Organisational governance starts with a set of ethical values that steer the behaviour of developers and managers towards the good of society [30, 31]. Of course, resistance and pressures to work fast may challenge the adoption of these values in practice [32]. Hence, these values need to be socialised with employees and stakeholders to create common goals and mental models, and supported by accountability mechanisms [29, 32].

Concerning the definition of use case's requirements, scholars stress stringent conditions for developing AI. Kahneman and Klein [33], for example, argue that it requires: a) specifying reliable success criteria, b) understanding similar cases and c) analysing conditions that render cost-effective algorithmic decision-making. Scholars support and expand these conditions by studying the empirical challenges of designing ethical AI. Regarding (a), they recommend success criteria to be defined

before any software development [34]. For one, data scientists and managers may be tempted to ex-post rationalise AI's performance or drift use cases for pressures to complete projects quickly [21, 25, 35, 36]. Regarding (b), scholars suggest reviewing existing systems in place, both internally and externally, to establish baseline metrics and assess the AI's feasibility [37]. Regarding (c), studies document that the cost-effectiveness analysis must extend beyond economic, technical and legal evaluations, and include ethical and environmental assessments. For instance, Taddeo et al. [38] recommend assessing the AI's carbon footprint, as some AI systems consume enormous computational power. Similarly, the AI should be assessed against the organisation's ethical values defined by organisational governance.

Item		Supporting information	Target respondent
Organisational Governance			
1.	The organisation has defined the set of values that should guide the development of AI systems	Description of the norms and values	Top manager responsible for AI
2.	These values have been published/ communicated externally	Short description of how values were communicated externally	Top manager responsible for AI
3.	These values have been communicated to internal AI project stakeholders	Short description of how values were communicated internally	Top manager responsible for AI
4.	A governance framework for AI projects has been defined	Short description of the AI governance frame-work, i.e., how adherence to the organisational values will be ensured and demonstrated in practice	Top manager responsible for AI
5.	The responsibility for ensuring and demonstrating that AI systems adhere to defined organisational values has been assigned	Name(s) of the person assigned	Top manager responsible for AI

Use Case			
6.	The objectives of the AI application have been defined and documented	Short description of the objectives of the AI application	Project manager
7.	The AI application has been assessed against the ethical values	Ethical assessment	Project manager
8.	Performance criteria for the AI application have been defined	Requirement specification document	Project manager
9.	The overall environmental impact for this AI application has been assessed	Assessment of the environmental impact of the AI application	Project manager

Table 2: Review items for the design stage

4.2 Stage 2: Development

Development is the core stage in the AI life cycle, as it sets the reference points for the model performance. In essence, developers leverage historical data to train an algorithm to make predictions [39]. Inappropriate or incomplete development processes may lead to epistemic concerns like inconclusive, inscrutable and misguided evidence [40–42], which challenge the validity of algorithmic predictions. The 'garbage in, garbage out' aphorism illustrates the problem: the input of distorted data, which misrepresent reality, leads to unreliable AI predictions as the output.

From a process perspective, these failures emerge when either the input (data and pre-processing) or the conversion (training the algorithm) are defective. Our analysis suggests two steps to address this problem: prepare and train. The prepare step concerns collecting the 'right' or 'good quality' data and transforming it with appropriate methods to ensure quality and compliance. Data quality covers criteria such as uniqueness, accuracy, consistency, completeness, timeliness and currency [43]. Agrawal et al. [44] argue 'the better the data, the better the prediction, the better the decision, the better the outcome'. After all, the statistical learning used by algorithms requires large datasets with appropriate attributes to make the correct inferences [45].

The training step concerns all the tasks for ensuring the model produces reliable predictions. It includes tasks such as selecting features, training, validating and tuning the model. Tuning ensures that the algorithm is trained to perform its best; it uses all the available information to reduce uncertainty in the outcomes. This is an iterative process, and model versioning is suggested to explain differences in model performance and compare models (e.g., through A/B testing) [39].

Question	Supporting information	Target respondent
Data		
10. The data used to develop the AI application has been documented	List of data used in the AI application	Project manager
11. Data used in the development has been checked for representativeness, relevance, accuracy, traceability (e.g., external data) and completeness	Data impact assessment; see e.g., IAF *Ethical Data Impact Assessment* or CNIL *Privacy Impact Assessment*	Project manager
12. The risks identified in the data impact assessment have been considered and addressed	Handling missing data; handling imbalance data; scaling; normalisation	Project manager
13. Legal compliance with respect to data protection has been assessed, e.g., GDPR	Data compliance assessment, including a list of protected attributes	Project manager
Model		
14. The source of the model has been documented	Source of the model	Project manager
15. The selection of the model has been assessed with regard to fairness, explainability and robustness	List of risks identified	Project manager
16. The risks identified in the model have been considered and addressed	List of assurance countermeasures	Project manager

Table 3: *Internal review questions for the development stage: 'Prepare' step*

Item	Supporting information	Target respondent
17. The strategy for validating the model has been defined	Brief description of the validation strategy	Project manager
18. The organisation documented the AI performance in the *training* environment	Performance on the training set in relation to agreed objectives	Data scientist
19. The setting of hyperparameters has been documented	Justification for the selection and levels of hyperparameters used	Data scientist
20. The model fulfils the established performance criteria levels	Documentation of model performance	Project manager

Table 4: *Internal review questions for the Development stage: 'Train' step*

4.3 Stage 3: Evaluation

During the *evaluation* stage, AI systems performance across different relevant dimensions are tested, measured, and assessed before they can be brought to the market [21, 25, 39]. Compared with traditional software development, AI projects require a dedicated evaluation stage in the life cycle. Notably, two steps are needed: *test* and *deploy.* The *test* step aims to assess how the AI system performs on unseen data across a set of dimensions, such as technical robustness, and adherence to ethical norms and values. To that end, organisations should instrument AI to measure performance. In turn, these instruments are used by developers to decide on when and how to refine the model [21]. Quantitative metrics alone are insufficient to assess AI systems. Therefore, developers should act as complementarities in reducing errors and biases, especially regarding input incompleteness [46].

The *deploy* step ultimately concerns deploying a tested model into the production environment. To arrive at that point, data scientists first need to define the serving strategy and its impact on users' privacy and security. Adopting pilots, such as canary deployment, minimise the risks of unforeseen failures.

Item	Supporting information	Target respondent
21. The strategy for testing the model has been defined	Short description of the validation strategy	Project manager
22. The organisation has documented the AI performance in the testing environment	Documentation model performance on the testing set in statistical terms	Data scientist
23. The model has been tested for performance on extreme values and protected attributes	Short description of performance on extreme values and protected attributes	Data scientist
24. Patterns of failure have been identified	FMEA, e.g., error curves, overfitting analysis, exploration of incorrect predictions	Data scientist
25. Key failure modes have been addressed	Short description of how to resolve or account for key failure modes	Data scientist
26. The model fulfils the established performance criteria levels	Documentation of model performance	Project manager

Table 5: Internal review items for the evaluation stage: 'Test' step

Item	Supporting information	Target respondent
27. The deployment strategy has been documented	Short description of the deployment strategy	Product owner
28. The serving strategy has been documented	Short description of the serving strategy	Product owner
29. The risks associated with the given serving and deployment strategies have been identified	Short description of identified risks	Product owner
30. The risks associated with the given serving and deployment strategies have been addressed	Short description of how to resolve or account for key risks	Product owner
31. The model fulfils the established performance criteria levels in the production environment	Performance in the production environment	Product owner

Table 6: Internal review items for the Evaluation stage: 'Deploy' step

4.4 Stage 4: Operation

The fourth principle of process theory states that unmanaged processes will deteriorate over time [20]. In context, this implies that even AI that performs well at launch will gradually decay [34]. This can lead to robustness and ethical failures. Most practitioners discount the importance of the actual operation of AI systems. However, research suggests that it is an essential and costly aspect of the AI life cycle [47]. Our analysis identifies two steps in the operation stage, *sustain* and *maintain*, which prevent common failures. *Sustain* refers to all activities that keep the system working, such as monitoring its performance, and establishing feedback collection mechanisms. As users interact with the AI system, they might use it in ways that were unforeseen by the developers, producing errors that need to be resolved [37]. *Maintain* refers to providing updates to keep the system running in good condition or improve it. This step involves defining regular update cycles [37] and establishing problem-to-resolution processes.

Item	Supporting information	Target respondent
32. The risks associated with changing data quality and potential data drift have been identified	A short description of the risks associated with data quality is captured (e.g., data drift, bias drift, feature attribution drift)	Product owner
33. The risks associated with model decay have been identified	A short description of the risks associated with model decay is captured	Product owner
34. The strategy for monitoring and addressing risks associated with data quality and drift; and model decay has been defined	Outline of monitoring strategy (e.g., error classification, critical threshold values for data drift and model decay)	Product owner
35. Periodic reviews of the AI applications with regard to the ethical values have been set	Review schedule and format	Top manager responsible for AI

Table 7: Internal review items for the operation stage: 'Sustain' step

Item	Supporting information	Target respondent
36. The organisation has a strategy for how to update the AI application continuously	Frequency of updates and documentation of model changes	Product owner
37. A complaints process has been established for users of the AI system to raise concerns or suggest improvements	Short description of the complaints process (e.g., point of contact)	Product owner
38. A problem-to-resolution process has been defined	Outline of problem-to-resolution process	Product owner

Table 8: Internal review questions for the Operation stage: 'Maintain' step

4.5 Stage 5: Retirement

This stage begins when organisations decide to take an AI system out of service, and ends when all elements have been disposed of adequately, archived or deactivated [24]. Our analysis suggests one step, *deactivate,* which includes all the activities from assessing the risks of deactivating an AI, to evaluating how to handle data records based on critical disposal needs that are specified in the agreements or the risk assessment.

Item	Supporting information	Target respondent
39. The risks of decommissioning the AI system have been assessed	Documentation of decommissioning risks	Product owner
40. The strategy for addressing risks associated with decommissioning the AI system	Outline of the strategy to manage the risks of decommissioning AI (e.g., data residuals: what will happen to data records, model accessibility and interfaces to other systems)	Top manager responsible for AI

Table 9: Internal review questions for the retirement stage

5 Summary datasheet

The requirements for the SDS are outlined in Annex VIII to the AIA, which states which information needs to be submitted upon the registration of high-risk AI systems in accordance with Article 51:

1. Name, address and contact details of the provider.
2. Where another person carries out submission of information on behalf of the provider, the name, address and contact details of that person.
3. Name, address and contact details of the authorised representative, where applicable.
4. AI system trade name and any ambiguous reference allowing identification and traceability of the AI system.
5. Description of the intended purpose of the AI system.
6. Status of the AI system (on the market, or in service; not placed on the market/in service, recalled).
7. Type, number and expiry date of the certificate issued by the notified body and the name of identification number of that notified body (where applicable).[5]
8. A scanned copy of the certificate referred to in point 7 (where applicable).
9. Member States in which the AI system is or has been placed on the market, put into service or made available in the Union.
10. A copy of the EU declaration of conformity referred to in Article 48.
11. Electronic instructions for use; this information shall not be provided for high-risk AI systems in the areas of law enforcement and migration, asylum and border control management referred to in Annex III, points 1, 6 and 7.
12. URL for additional information (optional). Providing this link is optional, yet in our view it is useful to include it here as well as in the *external scorecard*, which we are proposing below as an additional document to be made available publicly.

[5] This requirement applies to conformity assessments carried out by a third-party, the 'notified body', which yields a certificate valid for five years.

6 External scorecard

The ESC is a summary or overview document to be made available externally. It is a 'health check' to show the application of good practice and conscious management of ethical issues across the AI life cycle. It does not disclose any competitive or sensitive information, yet it describes the purpose of the AI system and provides an overview of key aspects of the ethical values behind the development of the AI system.

6.1 Content of the external scorecard

The ESC consists of four elements: purpose, values, data and governance. Like a balanced scorecard [48, 49], it covers retrospective, current and forward-looking aspects of the AI system. Conceptually, it is closely related to a 'model card' [50]. The elements can be freely chosen according to specific circumstances, yet we propose these four elements as the most meaningful aspects to be made available to customers and counterparties:

Item	Action
1. Purpose	Describe the AI system in terms of its objective and functionality.
2. Values	Outline the organisational values and norms that underpin the development of the AI system.
3. Data	A. Define the data used in terms of its public, proprietary and/or private nature. B. State whether the data used is internal and/or provided by a third party. C. Specify how consent has been secured for the use of this data. D. State whether the AI system uses protected attributes.
4. Governance	A. State the person responsible for the AI system. B. Provide a point of contact for any complaints or concerns. C. State the date when the initial AI system was deployed. D. Specify the dates of the last and next review of the AI system.

Table 10: The four aspects covered in the External Scorecard

The four quadrants of the external scorecard provide a qualitative overview of the AI system's functionality, underpinning values, data and governance. We suggest complementing this qualitative statement with a quantitative risk score. Conjointly, the qualitative and quantitative parts provide a more meaningful assessment.

6.2 Graphical representation of the external scorecard

The following picture illustrates a graphical representation of how the external scorecard could be made available to stakeholders:

Figure 7: Example of an external scorecard

PART II – THE REFERENCE

7 Defining the AI process flow

7.1 Defining AI development and operation as a process

This section defines AI systems as a process flow rather than a 'black box', and suggests that such a process is different from traditional software development. Specifically, we propose a set of life cycle stages and good practices to avoid two discrete AI failure modes: omission of critical tasks and incomplete application of critical tasks. For one, some organisations and data scientists responsible for AI are tempted to overlook proper development due to the pressures of delivering new products and services as fast as possible. The trouble with that approach is that AI is very susceptible to technical debt [47] – a cost that emerges from rework through prioritising speed over quality [51], and so it is fragile and prone to fail.

To define the proposed AI process flow, we reviewed the literature on software development and ML Ops to distil the stages for building AI systems. We conducted a comprehensive search in Web of Science, Google Scholar and Google search. We focused our analysis on software development life cycles documented in widely adopted standards such as ISO, IEC and IEEE standards. These standards suggest four life cycle stages in traditional software development: Concept, Development, Operation & Maintenance, and Retirement (e.g., ISO/IEC TR 24748-1 and ISO/IEC/IEEE 12207:2017). These stages are not prescriptive, but we maintained them as a starting point for consistency. So, we kept the same names to convey the scope and purpose of a set of development tasks. One caveat, of course, is that such standards were designed for traditional software development and not for the development of AI. Thus, following the standard recommendations, we proceeded to tailor the life cycle to the specific idiosyncrasies of AI.

Subsequently, we identified the differences between traditional software and AI development to refine the AI life cycle. Two salient differences were noted. First, machine learning outcomes result from statistical inference rather than ground truth. Second, once deployed, programmers spend less time monitoring, tracking changes and updating the model, as their efforts are put into automated, reproducible pipelines that take care of most of the updates when new data is available [39]. These differences suggest that AI developers are more prone to creating, or propagating, existing societal biases, albeit inadvertently. Accordingly, exhaustive model testing ex-ante the deployment stage is core to identifying and measuring any biases embedded in the AI system. Goodfellow et al. [21], for instance, suggest instrumenting the machine learning training and testing process to discover problems in performance. Similarly, others recommend an in-depth analysis of the trained model using multiple metrics [39]. In our context, we consider performance metrics and fairness metrics.

Addressing the need for rigorous testing, we added an evaluation stage to the traditional software life cycle model, resulting in five stages: Concept, Development, Evaluation, Operation and Retirement.

To analyse what is happening at each stage and how they can lead to ethical failure, each stage was divided into steps and each step was further subdivided into tasks. After reviewing different sources of machine learning life cycles and pipelines, we named the steps as we did for the stages: names convey the general purpose and scope of each step, and, when possible, we maintain the literature names. The final AI life cycle is composed of stages, divided into steps, subsequently divided into tasks. The result of our analysis is a set of tasks that providers perform at each stage of the AI life cycle, so as to ensure and show adherence to the requirements defined in the AIA (Table 11).

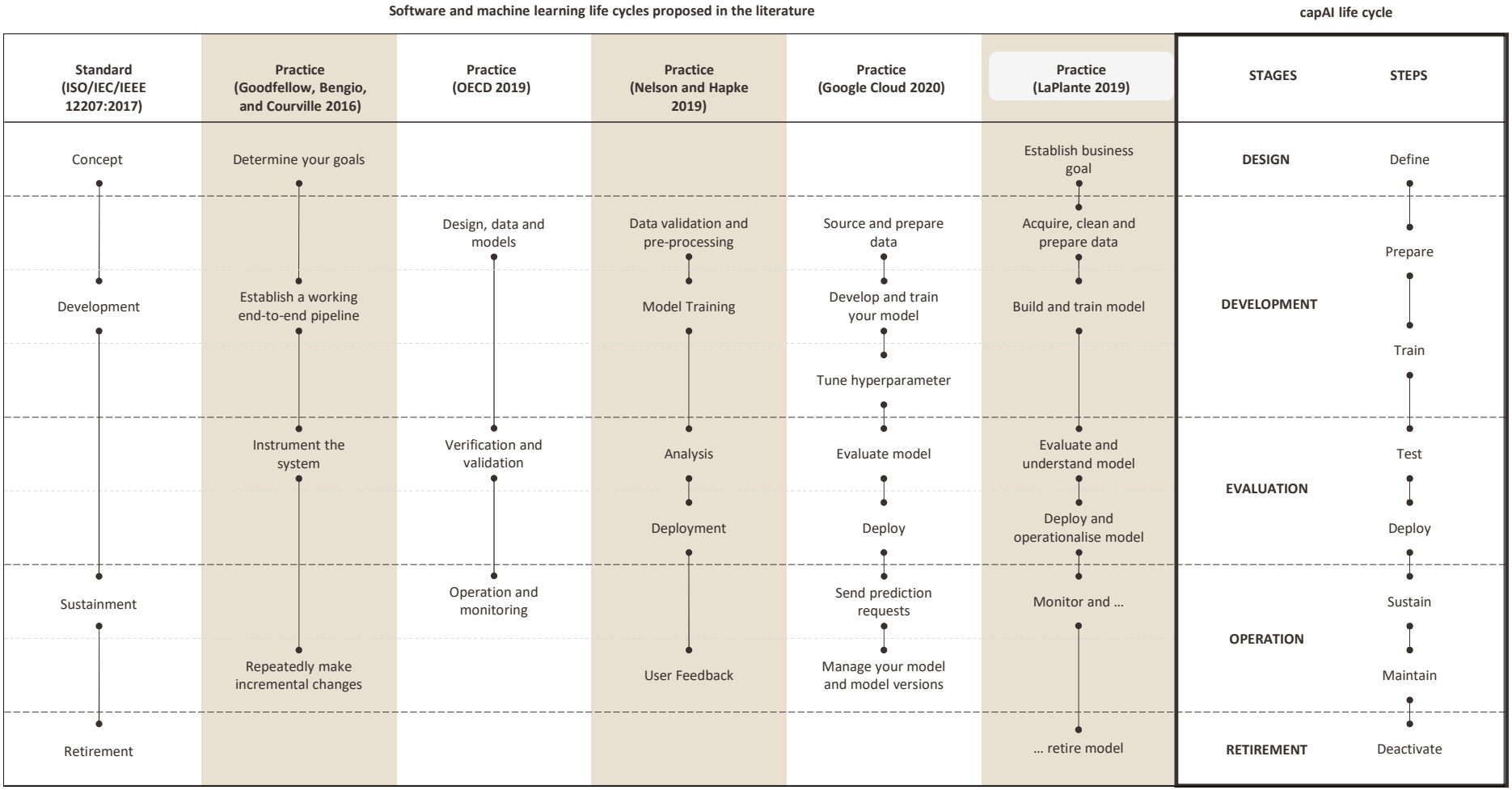

Table 11: Examples of AI and software life cycles coded into AI development steps and stages

7.2 The five key stages in the AI life cycle

We define five key stages in the life cycle of an AI system, from concept to retirement. These stages are enacted within three 'environments': training, testing and production environments. The term 'training environment' refers to an environment composed of the training and validation datasets. Similarly, the 'testing environment' refers to a separate environment composed of a testing dataset. Finally, 'production environment' refers to the setting where the deployed model makes predictions on new data it has not seen before, and is in actual operation for its intended purpose. In the following sections, we will discuss each stage in detail.

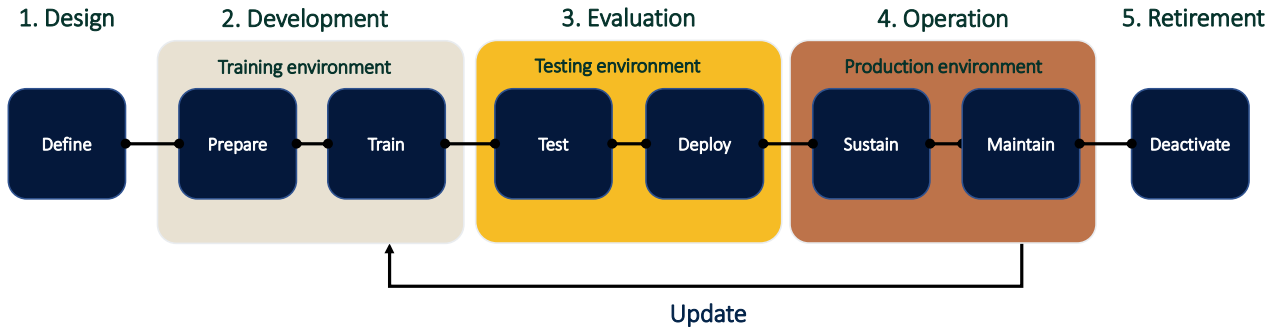

Figure 8: The five stages of the AI life cycle

7.3 The design stage

At the concept stage of the AI life cycle, organisations focus on specifying the data, model and variables to be used by an AI system. A specific problem that emerges is 'problem misspecification' – one of the key types of robustness, compliance and ethical failures of AI systems. It is defined as a functional form of the problem not being reflective of the true underlying problem [23]. Thus, an ethical and robust design of ML systems requires more than just a sound understanding of development techniques and algorithms; it requires defining the problem to solve its use case, risks, benefits and metrics to measure success/failure. To address this issue, the first step in the **capAI** life cycle focuses on tasks that help to minimise misspecification failures.

Formulate a use case

When developing an AI system, it is essential to first define the use case for which it is intended. This entails having an understanding of who the project stakeholders are, delineating a problem the AI system should help solve, and specifying how it does so. Defining a use case helps providers to clarify the goals of the AI systems and the requirements that need to be met for it to be beneficial, cost-effective and attractive for customers, as well as identifying its limitations and trade-offs. Among other things,

this requires identifying what existing solutions look like [25]. For example, before developing an AI system for credit card fraud detection, the provider should map internal systems in place, such as a rule-based algorithm, to discover how the AI use case will fit with existing organisational processes, its interdependencies and the minimum target value to beat for ML to be beneficial. Similarly, identifying external solutions, such as Bayesian Network Classifiers, helps with understanding the state of the art of the available solutions, supports error rate benchmark definition, and facilitates the selection of AI algorithms to test (more details on these are given in subsequent sections).

Defining an AI use case also concerns specifying the appropriate type of task and architecture [25]. Continuing with the credit card fraud detection example, the identified ML task should be a binary classification algorithm [25]. Then, the question is, does it require deep neural networks, or does it suffice with a simple classification algorithm? Deep neural network architectures are suggested particularly for use cases that fall into an 'AI-complete' problem,[6] like computer vision and natural language understanding [21]. By subjecting the use case to internal debate about the type of task and architecture, data scientists minimise AI failures caused by model misspecification.

Multiple ways of defining typical AI tasks are possible, as illustrated in Table 12 below.

Algorithm	Tasks	Use case examples
Supervised Learning	Regression	Market Forecasting Advertising Popularity Prediction
	Classification	Identity Fraud Detection Image Classification
Unsupervised Learning	Clustering	Customer Segmentation Recommender Systems
	Dimensionality Reduction	Meaningful Data Compression Structure Discovery
Reinforcement Learning		Skill Acquisition Game AI

Table 12: Examples of typical Machine Learning tasks

[6] 'AI-complete' denotes category of problems/subproblems in AI that indicates its complexity and presupposes its solution involves AI, as it cannot be solved with traditional programming techniques.

Assess the fit between AI system and problem type

This task requires evaluating if a specific AI system is the best possible solution for the use case's problem. AI use cases usually fall into one of the following project archetypes: improving, augmenting or automating an existing process. The selection of the archetype that best fits the AI system indicates the potential trade-offs between AI impact and feasibility (see Table 13 for a typology). For example, if the system aims to automate a process fully, it may indicate that an AI system has a high impact on service but low feasibility, due to complexity and uncertainty. Inherently, this archetype of projects involves bigger functional, security, reputation, legal and ethical risks that need to be acknowledged and addressed.

Typology	Description	Examples
Archetype 1	Improve an existing process	Improve code completion in an IDE
Archetype 2	Augment a manual process	Turning sketches into slides Help radiologists in reading images
Archetype 3	Automate a manual process	Fully self-driving car Automating customer service

Table 13: A typology of ML projects. Source: adapted from Full Stack Deep Learning

Translate use case into goals and metrics

Prior to any development activity, it is essential to agree on what the AI system aims to deliver and how to measure success to prevent post-hoc performance rationalisation. Use case drift can be prevented by translating the use case into error metrics and targeting values to measure success during the concept stage [21, 25, 52]. Unfortunately, no silver bullet exists to tackle the complex measurement problem of AI performance. Appropriate measures are required to prevent target variable misspecification. Organisations need to define metrics that assess different performance dimensions in the use case context. **capAI** considers at least two dimensions: robustness and ethical performance.

Robustness error metrics refer to those used to measure AI prediction capabilities, such as accuracy and specificity, and its applicability depends on the use case specifications (see Tables 14 and 15 for examples). For example, accuracy (one of the most common metrics in classification tasks) is a poor metric to characterise AI's performance on rare events identification problems (e.g., a rare disease) and, more broadly, in unbalanced-classification problems [21]. In such scenarios, using a combination of precision and recall is appropriate. The main point, however, it that

understanding the organisational metrics already in place facilitates the selection of robustness metrics. After all, the AI should perform better than the existing system. Nevertheless, many performance metrics are possible here, and it may be appropriate to define new metrics to evaluate a specific AI system. For examples of common metrics see Tables 14 and 15.

The second family of metrics involves those used to ensure model fairness [23]. Examples include equalised odds, Theil index and demographic parity (for more examples, see Table 20). The target value of those metrics should be identified using industry benchmarks to ensure the AI application is safe, fair, cost-effective and appealing to customers.

Metric	Definition	Possible Applications
Mean Squared Error (MSE)	$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$	Used in applications where penalising larger errors is important. For example, in a house price prediction task based on the number of rooms.
Mean Absolute Error (MAE)	$\frac{1}{n}\sum_{i=1}^{n}\lvert Y_i - \widehat{Y}_i\rvert$	Used in regressions that consider only the absolute error distance.

Table 14: Examples of typical robustness metrics for regression tasks (e.g., Linear Regression, Decision Tree Regression, Random Forest Regression). Here, $\boldsymbol{n}$ is the number of data points; $\boldsymbol{Y_i}$ and $\boldsymbol{\widehat{Y_i}}$ are the actual and predicted target value at point $\boldsymbol{i}$

Metric	Definition	Possible Applications
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	Used in classification tasks where an approximately equal number of samples belong to each class.
Precision	$\frac{TP}{TP + FP}$	Used when the cost of False Positive is high. For example, in email spam detection, misclassifying an email as spam (False Positive) might lead to the user losing valuable information.
Recall or Sensitivity	$\frac{TP}{TP + FN}$	Used when the cost of False Negative is high. For example, in fraud detection, misclassifying a fraudulent transaction as non-fraudulent (False Negative) may be expensive for a bank.
Specificity	$\frac{TN}{TN + FP}$	Used to determine a model's ability to predict if an observation does not belong to a specific category.
F-score	$\frac{2TP}{2TP + FP + FN}$	F-score is needed when a balance between Precision and Recall is sought.

Table 15: *Examples of typical robustness metrics for classification tasks. Here, TP stand for the True Positives, TN for True Negatives, FP for False Positives and FN for False Negatives*

Translate use case into data needs

This step concerns identifying the data needed for an AI system to solve the use case's problem successfully. As part of this step, organisations define what data is used both as predictive features and as targets [23]. To facilitate data sourcing and avoid problem misspecification, project managers should ensure that the data captures accurately the problem and avoids proxy variables. For example, if the goal is 'to predict a defendant's risk to public safety – as most risk assessment tools are – the objective must be whether a defendant is likely to commit an offence that justifies pretrial detention, not whether the defendant is likely to be arrested or convicted of any offence in the future' [53]. Similarly, data scientists should ensure that predictive features represent the underlying problem. One way to do so is by checking that data meets six criteria: uniqueness, accuracy, consistency, completeness, timeliness and currency [43].

Cost-benefit analysis

It is good practice for organisations to run a cost-benefit analysis of an AI system to evaluate the trade-offs of algorithmic decisions (Table 16). For instance, if the AI system aims to improve an existing process, the organisation should question whether and by how much the model will improve performance. As obvious as it may seem, a cost-benefit analysis should include an estimation of the costs of developing AI systems, such as sourcing and labelling data.

Typology	Description	Examples
Archetype 1	Improve an existing process	Do the models improve performance? Does performance improvement generate business value? Does performance improvement lead to a data flywheel?
Archetype 2	Augment a manual process	How good does the system need to be to qualify as useful? How can enough data be collected to make it that good?
Archetype 3	Automate a manual process	What is an acceptable failure rate for the system? How can it be guaranteed that it will not exceed that failure rate? How can data from the system be labelled inexpensively?

Table 16: *Key cost-benefit analysis questions to consider by AI project archetype. Source: Full Stack Deep Learning 2020*

Risk analysis

Before moving forward to the development stage, organisations should conduct a risk analysis of the AI use case from at least three perspectives: operational risk of failure, security risk of interference, and legal risks arising from failure. Although it might be tempting to compartmentalise these risk, such an approach would silo the information, inhibit the apprehension of risks interaction, and disperse responsibility [54]. Thus, an active and integrative risk assessment is preferable. The risk analysis could take many forms; below are some suggestions:

- ***Operational risks:*** This involves assessing risks of the AI system failing while in operation, using a Failure Mode Effect Analysis (FMEA) or similar. The assessment should include assessing whether such failure can affect negatively fundamental human rights [55]. For example, in a use case involving AI in loan management, the retail bank must assess whether it can lead to

exclusive access or use by certain groups (including age, gender, ethnic background and disability) [56].

- ***Security risks:*** Some common security risks to consider in this discussion relate to evasion, sabotage and privacy attacks. Evasion attacks are those where hackers attempt to trick the AI to make false predictions. Data poisoning refers to attacks where data is corrupted to make the model learn the wrong inferences and compromise the prediction's integrity and availability. Finally, data privacy and confidentiality risks concern retrieving sensitive data from the model. Discussing security risks should also contemplate where data is stored (e.g., on-device, own servers or on a cloud server).

- ***Legal risks:*** Once the use case has been detailed, the use case should be submitted to the legal team for review. The legal team should ensure the proposal meets general requirements such as GDPR/CCPA, and context-specific regulations, such as SOX compliance in financial applications. Furthermore, the legal team should assess compliance with product safety and liability regulations (forthcoming changes on the EU Product Liability Directive, and the EU General Product Safety Directive). As governments keep updating and developing regulations to protect citizens and foster digital innovation, monitoring future regulatory changes and discussing the potential impact with the project stakeholders is essential.

Once the proposal has been approved, the project is ready to go into the development stage.

7.4 The development stage

Sourcing the data

Data is the key ingredient for developing ML systems. Without the right amount and quality of data, an AI system will be doomed to fail – and even if not, it may unintentionally spread established biases in society. The task of sourcing data starts with extracting and combining records from multiple data sources into a single dataset – an activity sometimes referred to as 'extract, transform and load' (ETL). Organisations rarely own all the data required to train an AI model, so they may start an additional data collection process. This process involves deciding whether data will be collected in-house, for example, through setting up sensors or different data vendors. In any case, the collected data must comply with the relevant regulations (e.g., GDPR, CCPA), and permission should be obtained for the specific use case.

Of course, compliant data is not enough for building ethical AI systems. Low-quality data could be equally pervasive, and can also lead an AI system to learn bias and propagate socially derived artefacts that disadvantage particular groups [56]. In this context, data quality is determined by six properties: uniqueness, accuracy, consistency, completeness, timeliness and currency [43]. In the case of supervised

learning, this step further includes specifying the data labelling approach [57, 58]. For example, an organisation that crowdsources the image labelling should summarise labellers, including geographical diversity (e.g., number of human labellers, nationality).

It is important to control biases that can lead to an inaccurate representation of the problem or propagate socially derived artefacts that disadvantage particular groups [56]. For example, reporting bias [59, 60], selection bias [61] and group attribution bias [23] (see Table 17 below). A final consideration involves versioning the data [39]. As data is added constantly, it is key to version any data included.

Description	Bias	Manifestation	Definition	Example
Discrimination in data	Reporting and measurement bias		'Reporting bias arises from how we choose, utilize and measure particular features.' [61]. An example of this bias is when you notice and report atypical situations, ignoring ordinary characteristics. In the context of image labelling, reporting bias may creep in from labellers' tendency to document 'what is worth saying' instead of 'what is in the image'.[60]	A commonly cited example of reporting bias is the one encoded in COMPAS, a recidivism risk prediction tool that used the number of personal and family arrests as a proxy variable for the risk of committing a crime. As minority groups are policed more often, they have higher arrest rates. Yet it would be a mistake to conclude that minority groups represent a greater danger to society, as they are monitored and controlled differently from other groups. [61, 62]
	Group attribution bias	In-group bias	In-group bias arises when you favour members of a group to which you belong or those who exhibit characteristics similar to yours. [63]	'Two engineers training a résumé-screening model for software developers are predisposed to believe that applicants who attended the same computer-science academy as they both did are more qualified for the role.' [64]
		Out-group homogeneity bias	Conversely, out-group bias arises when you stereotype members of a group or assume their characteristics as more uniform.[65]	'Two engineers training a résumé-screening model for software developers are predisposed to believe that all applicants who did not attend a computer-science academy do not have sufficient expertise for the role.' [64]
	Historical bias		Historical bias is the existing bias in the world; 'traditional prejudices that are endemic in reality.' [66] This issue may creep into the model from the collected data even under perfect sampling. [62]	An example of historical bias 'can be found in a 2018 image search result where searching for women CEOs ultimately resulted in fewer female CEO images due to the fact that only 5% of Fortune 500 CEOs were women – which would cause the search results to be biased towards male CEOs These search results were of course reflecting the reality, but whether or not the search algorithms should reflect this reality is an issue worth considering.' [62]
	Omitted variable bias		'Omitted variable bias occurs when one or more important variables are left out of the model.'[62, 67–69]	An example for this case would be a model to predict, with relatively high accuracy, the annual percentage rate at which customers will stop subscribing to a service. But someone soon observes that the majority of users are cancelling their subscription without receiving

				any warning from the designed model. Now imagine that the reason for cancelling the subscriptions is appearance of a new strong competitor in the market that offers the same solution, but for half the price. The competitor's appearance was something that the model was not ready for; therefore, it is considered to be an omitted variable. [61]
Non-representative data	Selection bias	Coverage bias	The population represented in the dataset does not match the population that the machine learning model is making predictions about.	Consider a model trained to predict people's emotions from their facial expressions. Coverage bias may arise if you train the model using European face images and deploy it to predict Asians' and Africans' emotions, as they may express some emotions with different expressions.
		Participation bias/non-response bias	Users from specific groups opt out of surveys at different rates than users from other groups [64]	A model is trained to predict future sales of a new product based on phone surveys conducted with a sample of consumers who bought the product and with a sample of consumers who bought a competing product. Instead of randomly targeting consumers, the surveyor chose the first 200 consumers that responded to an email, who might have been more enthusiastic about the product than average purchasers. [64]
		Sampling bias	This bias arises when the data collected to make inferences does not represent a random sampling of the subgroups. Consequently, the model inferences may not generalise to all subgroups. In practice, this bias stems from other biases, such as self-selection bias, exclusion bias and preferential sampling. [61]	This bias can be observed in opinion polls, where more enthusiastic people are more likely to complete the poll. [61]

Table 17: Common data issues that can affect the integrity of sourced and labelled data. Based on [23], the table is expanded to include definitions and examples

Analysing the data

Once data is sourced, the next task is to analyse and understand the data that will be used to train an ML system – including its distribution and integrity [70–72]. Multiple data validity tests should be carried out to check input biases and ensure that the data represents the underlying use case. Examples include verifying that statistics show expected distributions and ranges; identifying and exploring outliers; testing for adherence to data schema, range constraints and meta-level requirements; verifying that privacy controls are in place; and using cross-validation, among others [34, 39]. It is further recommended that data scientists annotate their assumptions and compare them with actual data statistics to prevent anchoring bias [34].

Following this, data scientists should select a subset of predictor variables to train the AI. In some cases, the use of race, gender, colour and other variables (see Table 18) may result in discrimination against minority groups. To ensure equal treatment among actors with protected attributes, organisations must design appropriate mechanisms to prevent developers from drawing on such features [34]. Other variables may act as proxies for protected attributes, which may lead to unintentional discrimination among classes. To avoid this issue, data scientists should test and document any correlation between protected attributes and any other features [23]. Depending on the magnitude of the correlation, data scientists must decide whether to apply any pre-processing techniques to minimise the risk of generating unfair outcomes.

Table 18 presents some examples of protected attributes involving housing, credit applications, and working in the context of the US and Australia. In the case of the EU, a comprehensive list of protected attributes for housing is presented by Silver and Danielowski [73]. Depending on the use case, other appropriate legislation to consider is the following Directives: Directive 2000/43/EC, Directive 2000/78/EC, Directive 2006/54/EC, Directive 2004/113/EC.

Attribute	FHA	ECOA	FWA
Race	x	x	x
Colour	x	x	x
National origin	x	x	
Religion	x	x	x
Sex	x	x	x
Familial status	x		
Disability	x		x
Exercised rights under CCPA		x	
Marital status		x	x
Recipient of public assistance		x	
Age		x	x
Sexual preference			x
Family or carer responsibility			x
Pregnancy			x
Political opinion			x
National extraction or social origin			x

Table 18: *Examples of Protected Attributes in housing and credit in the US and Australia. Based on [61]. FHA stands for the Fair Housing Act, ECOA stands for the Equal Credit Opportunity Act and FWA stands for the Fair Working Act*

Preparing the data

Once data scientists have cleaned, validated and achieved an understanding of the data, they need to format it to best suit the training runs. Many techniques for preparing data are domain specific (such as natural language processing or image recognition) or type-specific (supervised, unsupervised, reinforcement learning) [25]. For example, in the case of supervised learning, labels might need to be converted into multi-hot vectors [39]. However,

it has been suggested that some data preparation techniques are good practice in all domains. These are vectorisation, value normalisation and handling missing values [25].

Vectorisation. This is a technique to speed up code execution by removing loops. A common vectorisation technique involves converting the data into tensors. Speeding up the code execution leads to performance benefits and reduces programs' energy consumption.

Normalisation. This is a technique to adjust the values of numeric features into a common scale, without distorting the data distribution or losing information. For example, a dataset contains two features, one with values ranging from 0 to 10 and the other ranging from 1M to 10M. Due to the magnitude difference between these features, the larger one will have higher weight on the final model. Normalisation prevents this by adjusting the features to a common scale, which normally involves transforming each feature independently with a mean of 0 and a standard deviation of 1 [25].

Handling missing values. Finally, almost every dataset will have some observations with missing values in one or multiple features, which need to be handled to prepare the data for training. Handling missing values might involve deciding whether to discard a feature, apply imputation techniques or even create a new variable to tell the model whether a categorical value is missing [25]. At first, the data scientist should establish if the missing values are distorting the picture of the true population by determining the missing data mechanism, i.e., whether the data is missing at random (MAR), missing completely at random (MCAR), or missing not at random (MNAR). The data mechanism dictates the approach to handle missing values. For example, if the data is not missing at random, adopting imputation techniques would be dangerous and would affect the integrity of the data by propagating biases. Multiple tests are possible to diagnose the missing data mechanism, but often they should be combined with an inspection of the data collection process [45]. In the case of the imbalanced classification task, it might be possible to apply data augmentation techniques such as Synthetic Minority Oversampling Technique or SMOTE for short [74, 75].

Splitting the data

As with all other statistical methods, ML needs to achieve generalisation over unseen data to be reliable and trustworthy. This requires constant evaluation before the model is released for the production environment. Thus, before engaging in training the model, data scientists should first specify the evaluation protocol used later to estimate the success of the model [25]. Usually, two types of ML evaluation are applied: *validation* and *testing*. Validation is used for estimating prediction error and model selection/tuning. Testing is used to assess the generalisation error of the final model [45]. The selection of the evaluation protocol dictates the way available data should be split. Three broad categories of validation are possible: hold-out validation, k-fold cross-validation and iterated k-fold validation, depending on the amount of data available (see Table 19). Testing is usually performed on a hold-out set, kept in a 'vault' and used only at the end of the data analysis to prevent data leakage [45]. To accomplish this goal, data scientists should split data into multiple sets depending on the evaluation protocol selected. If the data is big enough, it is possible to select the data into

training, validating and testing sets. Otherwise, other strategies are more appropriate. It is essential to avoid redundancy and keep the sets disjoint when splitting the data. Furthermore, in classification tasks, data scientists should shuffle or stratify the data to ensure data representativeness, i.e., that the subsets accurately reflect the characteristics of the larger group [25].

Evaluation Protocol		Description	Data Split
Validation	Testing		
Hold-out validation	Hold-out testing	Select this strategy if a large dataset is available	
K-fold cross-validation	Hold-out testing	Select this strategy if not enough samples for hold-out validation to be reliable	
Iterated K-fold validation	Hold-out testing	Select this strategy for accurate evaluation if only a small dataset is available	

Table 19: *Examples of evaluation protocol strategies and respective data split. Based on [25, 45]. Data split into the training set (train), validation set (V) and testing set (Tt). A common data split is 50% training, 25% validation and 25% testing, while many other alternatives are of course possible.*

Training the model

Once all the data has been prepared, it is time to train the model. The goal here is to establish a baseline from which to improve the system, usually by applying feature engineering techniques [21, 25]. A common rule of thumb for model training is the application of commonplace algorithms (for example, those applied for similar tasks) rather than incautiously applying an obscure algorithm [21]. Obscure algorithms catalyst AI epistemic failures as they may produce inscrutable and misguided evidence [41].

Feeding a model with carefully curated data is insufficient to produce ethical and robust AI; further considerations are needed in training and feature engineering decisions to prevent AI failures. For example, suppose the use case concerns a binary classification problem, and the model is trained on unbalanced data (there are more examples of one class). In that case, the resulting AI will produce models biased towards the dominant category [39]. Identifying these scenarios, and applying appropriate countermeasures, such as adjustments on the loss function, adopting SMOTE, or collecting more data, become essential to de-risk the AI implementation.

Occasionally, the use case error metrics cannot be directly optimised by an algorithm (e.g., ROC AUC), so appropriate adjustments are required. In classification tasks that use

ROC AUC (error metric), it will suffice to use cross-entropy (loss function) [25]. Failing to make a suitable translation between error metric and loss function might lead to model drift.

Other AI failures might be induced by the feature engineering practice itself, as it might embed experimenter bias into the AI model. In context, experimenter bias (or observer-expectancy effect) refers to instances where data scientists intentionally or unintentionally influence the model by selecting features that match their predisposed notions or beliefs [53]. To minimise the risk of these failures, data scientists should declare whether feature engineering techniques were applied, and the rationale behind the deletion or creation of new features.

Validating the model

Once a new model has been trained, data scientists should validate it by instrumenting the model to measure the loss and comparing predictions to targets in the training and validating datasets. Validating occurs in two forms: software validation and performance validation. Software validation refers to tasks undertaken to debug the model implementation and ensure that no software errors creep in during the development process. For example, if the loss of the training dataset is high, it might be due to model underfitting or software defects. Then, appropriate testing should be undertaken to establish and resolve the cause of this failure. In this case, fitting a tiny dataset and exploring the model behaviour will suffice [21].

Performance validation refers to the ability of the model to perform and generalise over data. This is usually done on a separate dataset, called the validation dataset. Nevertheless, other alternatives are possible. A simple observation of the validation loss indicates the generalisation ability and technical robustness of the model. Validating the model result on feedback is used to tune the model.

Tuning the model

Establishing a validated model that performs is essential but not necessarily sufficient for your use case. For one, the model might not generalise over other data. Thus, the key question is whether the proposed AI model is the best possible to come up with available data. Failing to tune the model is unethical. It is irresponsible and wasteful to underutilise the available data, as this may result in lower quality predictions and output disparities. Therefore, this step concerns adjusting the model configuration to get the ‘best’ possible AI model. In practice, validation and tuning are intertwined and happen in parallel. The basic approach involves modifying the model, training it and validating the results iteratively until achieving the best possible combination [25]. Some strategies include conducting further feature engineering, such as adding, removing and creating a new combination of features; applying regularisation, such as LASSO and RIDGE; collecting and annotating more data; and changing hyperparameters, such as the learning rate, number of layers and their size (for neural networks).

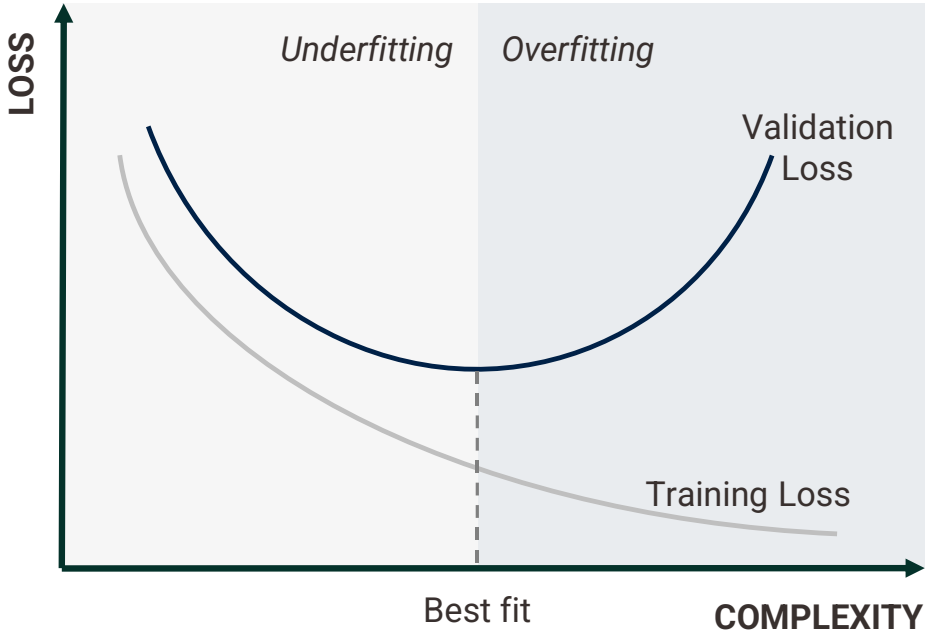

Figure 9: *Overfitting vs underfitting in model tuning*

The right model lies in the interface between overfitting and underfitting, i.e., where the model is capable of fitting the training data and at the same time generalising over unseen examples. The only way to determine where the best model fit lies is by crossing it [25]. That requires playing with different models and hyperparameters. For example, if the use case applies deep learning, some parameters can be modified by adding layers, increasing their size, and training for more epochs (in deep learning). It needs to test the model systematically by changing some parameters until the validation loss starts degrading. At this point, the model starts overfitting the data [25]. This configuration has statistical power and can be used in the tuning task.

Tuning hyperparameters can be done manually or automatically. Manual tuning an requires understanding of the relationship between hyperparameters, training error, generalisation error and computational resources, and thus, the approach should be decided wisely (for an overview of the effect of changing various hyperparameters please see [21]). On the other hand, automatic hyperparameter tuning approaches include Grid search, random search [76], and Bayesian optimisation [77–79]. Finally, it is worth noting the existence of automatic hyperparameter tuning services without opening the black box, such as Tune [80], Google Vizier [81], Auto-WEKA [82], Auto-SKLearn [83] and Mistique [84].

Once a final model has been tuned, organisations should record the training and validation loss, as well as the final hyperparameter configuration to ensure model traceability and replicability.

7.5 The evaluation stage

Testing for robustness

Once a model has been trained, validated and tuned, it is time to test it in the hold-out testing dataset and conduct a detailed analysis of the model performance. Using the metrics specified upfront, data analysis should calculate the testing error and compare it with that of the training dataset to diagnose any model issues. Some typical issues include overfitting/underfitting, bugs, validation issues and data issues. For example, suppose the testing set performance is much worse than the validation one. In that case, this might indicate that the validation procedure was not appropriate or that the model is overfitting the data [25].

To identify the root causes of performance problems, visualising the model in action is recommended in order to observe examples of how the model actually performs in a given task [21]. Visualisation is a valuable practice because it prevents organisations from falling prey to the automation bias that emerges from merely focusing on quantitative performance metrics and helps developers to identify and solve robustness threats. Consider, for instance, the development of an AI to recognise vehicles and pedestrians on the road for self-driving car applications. While developers may find it convenient to understand the model performance using only quantitative performance metrics (e.g., the accuracy of predicting a car), they may ignore key threats that lead to potentially fatal consequences. To prevent this issue, developers should pick a random sample of pictures with both cars and pedestrians and compare if the model predicted labels that match the objects on the picture. Developers should also apply a similar strategy to visualise examples that the model fails to model correctly. By conducting this visual inspection, organisations can identify systematic issues related to data collection, preparation and labelling [21] (e.g., if the model always fail to identify a certain type of vehicle) and hence, take corrective action.

Testing for discrimination

Another reason for conducting an in-depth analysis of the model performance is to discover and correct potential sources of discrimination that lead to unfair outcomes [23]. In order to choose from multiple strategies to avoid discrimination, developers first need to define and operationalise model ‘fairness’ based on the context and use case. Thus, it is essential to return to the values and norms, the use case, the identified protected attributes and its operationalised metrics, and have a discussion with different stakeholders to understand what fair outcomes should look like to define the appropriate correction approach. This goes beyond technical feasibility and requires adjoining regulation and ethics [53].

A final solution for testing for discrimination usually involves evaluating the outcomes of different fairness metrics and specifying an interval for a solution to be considered fair (examples of discrimination metrics are presented in Table 20). Furthermore, it may involve adopting post-processing techniques to correct for unfairness and minimise discriminatory outcomes [85, c.f. 86].

Metric	Definition	Reference
Statistical Parity Difference	The difference in the rate of favourable outcomes between the unprivileged group and the privileged group.	
Equal Opportunity Difference	The difference of true positive rates between the unprivileged and the unprivileged groups.	
Average Odds Difference	The average difference of false positive rate (False positives/ negatives) and true positive rate (true positives/positives) unprivileged and privileged groups.	
Disparate Impact	The ratio of the rate of a favourable outcome for the unprivileged group to that of the privileged group.	
Theil index	Measures the inequality in benefit allocation for individuals.	
Euclidean Distance	The average Euclidean distance between the samples from the two datasets.	
Mahalanobis Distance	The average Mahalanobis distance between the samples from the two datasets.	
Manhattan Distance	The average Manhattan distance between the samples from the two datasets.	
Equalised Odds	A predictor $\hat{Y}$ satisfies equalised odds with respect to protected attribute A and outcome Y, if $\hat{Y}$ and A are independent conditional on y. P ($\hat{Y} = 1 \mid A = 0, Y = y) = \hat{Y} = 1 \mid A = 1, Y = y), y \in (0,1)$ This means that the probability of a person in a negative class being incorrectly assigned a positive outcome should both be the same for the protected and unprotected (male and female) group members [123]. In other words, the equalised odds definition states that the protected and unprotected groups should have equal rates for true positives and false positives.	(Mehrabi et al. 2019)
Equal Opportunity	This means that the probability of a person in a positive class being assigned to a positive outcome should be equal for both protected and unprotected (female and male) group members. In other words, the equal opportunity definition states that the protected and unprotected groups should have true positive rates.	(Mehrabi et al. 2019)
Demographic Parity or Statistical Parity	A fairness metric is satisfied if the results of a model's classification are not dependent on a given sensitive attribute. For example, if both Lilliputians and Brobdingnagians apply to Glubbdubdrib University, demographic parity is achieved if the percentage of Lilliputians admitted is the same as the percentage of Brobdingnagians admitted, irrespective of whether one group is on average more qualified than the other.	https://developers.google.com/machine-learning/glossary/fairness
Fairness Through Awareness	An algorithm is fair if it gives similar predictions to similar individuals [43, 73]. In other words, any two individuals who are similar with respect to a similarity (inverse distance) metric defined for a particular task should receive a similar outcome.	(Mehrabi et al. 2019)

Fairness Through Unawareness	An algorithm is fair as long as any protected attributes are not explicitly used in the decision-making process [53, 73].	(Mehrabi et al. 2019)
Treatment Equality	Treatment equality is achieved when the ratio of false negatives and false positives is the same for both protected group categories.[14].	(Mehrabi et al. 2019)
Test Fairness	A score S = S(x) is test fair (well-calibrated) if it reflects the same likelihood of recidivism irrespective of the individual's group membership, R. That is, if for all values of s, P(Y =1|S=s,R=b)=P(Y =1|S=s,R=w) [31]. In other words, the test fairness definition states that for any predicted probability score S, people in both protected and unprotected (female and male) groups must have an equal probability of correctly belonging to the positive class.	(Mehrabi et al. 2019)
Counterfactual Fairness	A fairness metric that checks whether a classifier produces the same result for one individual as it does for another individual who is identical to the first, except with respect to one or more sensitive attributes. Evaluating a classifier for counterfactual fairness is one method for surfacing potential sources of bias in a model.	
Fairness in Relational Domains	A notion of fairness that is able to capture the relational structure in a domain – not only by considering attributes of individuals but by taking into account the social, organisational and other connections between individuals [44].	(Mehrabi et al. 2019)
Conditional Statistical Parity	For a set of legitimate factors L, predictor Y satisfies ^^conditional statistical parity if P(Y |L=1,A = 0) = P(Y|L=1,A = 1) [37]. Conditional statistical parity states that people in both protected and unprotected (female and male) groups should have an equal probability of being assigned to a positive outcome given a set of legitimate factors L.	(Mehrabi et al. 2019)

Table 20: *Examples of fairness metrics. Based on Mehrabi et al. [61].*

Refining the model

The final task in the evaluation stage involves refining the model. If the model has performed poorly either in the robustness, fairness or assessment of the ethical principles, data scientists should seek to refine it. Refinement involves making incremental changes, gathering and labelling more data, retraining the model, changing the algorithm, and/or tuning the hyperparameters based on the specific finding of the instrumentation available [21]. If there are no bugs in the implementation, the go-to strategy is gathering more relevant data to retrain the model, subject to availability and costs. Otherwise, data scientists should decide on other alternatives to improve performance. This task then requires continuous revaluation of the model until it meets the performance goals and ethical requirements. Once the model has passed all the tests and the error has been recorded, the organisation should deploy and maintain the system.

Instrument the model to capture model decay

Over time, models will degrade and generate errors that need to be addressed. This can happen for multiple reasons. For example, because the data used to train and test the model fails to represent the production environment, the model relevance is much lower than expected. Another common cause is the idea of concept drift: over time, the production data will change in unforeseen ways, and so the model predictions decay. Imagine the effect of COVID-19 on a model trained on pre-pandemic data. It will perform poorly now as the model cannot make appropriate statistical inferences. Even without a pandemic, model decay over time is highly dependent upon the use case and context. For example, in card fraud detection, the decay time is measured in days, while for the image search engine, in years [25]. Independent of the decay time, it is key to instrumenting the system and raises alarms when performance degrades, or something wrong happens, including writing tests for drift, outliers and downtime.

Pilot and test the model

Just as with any other software deployment, it is essential to conduct a pilot before releasing the model for full operation. This is the time to validate the system performance in a controlled production dataset. It is therefore possible to identify and correct the problem promptly, without serious consequence on the service. Data scientists usually adopt methods such as canary deployment, where the model is released for a small segment. Here, the model performance should be monitored, looking for regressions and integration issues.

Selecting a deployment strategy

Once a model that satisfies the design requirements is available, it is essential to convert it into a product that can be served to the customers. Some available options deploy the model as a REST API, on a device, and in the browser [25]. Deploying as a REST API is the most common option as the model runs predictions on demand on a web server. This can be done either by deploying the code to virtual machines, as containers using orchestration platforms (e.g., Docker and Kubernetes) [87], or as a ‘serverless function’ [88]. Deploying on a device, such as a smartphone or a microcontroller, is another useful alternative when limited internet connectivity halt model performance or data sensitivity concerns exist. In those cases, it is essential to verify that the device memory can cope with the model requirements. Finally, deploying in the browser resembles on-device implementation, only that, in this case, the model runs on the user CPU/GPU. Similarly, it is key to check the RAM constraints to ensure the model can serve effectively. In the last two cases, a copy of the model is stored in the local devices, and so one must ensure that the model does not involve any private information [25].

Rollout of the system

If the model has passed all the quality, integration, inclusion and ethical tests, it is ready to continue to the rollout, when it enters 'production', i.e., makes predictions on new datapoints.

7.6 The operation stage

Monitor the serving system

The work does not end when the AI system has been deployed. Using the instruments created upfront in the development stage, organisations must move into a monitoring stage. Good monitoring involves observing statistics, data distribution and business use change – users' interactions with the model predictions. Platforms provide monitoring solutions (e.g., Amazon SageMaker and Domino Data Lab). Monitoring should log any issues and record any actions taken to remediate them. This requirement is consistent with the AIA requirement for logging of key events.

Establish feedback mechanisms

Sustaining AI systems involves designing appropriate mechanisms for collecting customers' feedback and improving the model. Feedback can be collected either implicitly, for example, users' actions support product inferences [37], or explicitly, referring to instruments designed to let users intentionally provide feedback, for example, via surveys, Likert scales, likes or open text field.

By collecting feedback and monitoring, organisations explore the AI system's performance and identify errors and opportunities for improvement. However, it is easy to ignore these errors until a major incident occurs. Defining ex-ante what constitutes an error prevents this problem. An example of a broad error classification relies on distinguishing between errors due to users misusing the model ('user errors'), the system being too inflexible to meet user needs ('system errors'), or the system making erroneous assumptions about the user ('context errors') [37]. Users must be informed of this implicit feedback collection in the service terms.

Once errors are classified, it is time to identify their root causes. For example, one can classify errors by prediction and training errors, which occur when the available training data errors cap model performance; input errors, which occur when the user enters inputs that the model is unable to recognise; relevance errors, which occur when the model produces predictions unable to meet the customer's needs; and system hierarchy errors, which occur when the user connects the model with another system and the system is unable to recognise the hierarchical controls [37].

Define regular updates cycles

AI systems learn over time, i.e., they adapt to environmental changes and update their internal decision-making logic based on new input data. However, as mentioned earlier, many organisations fail to achieve this, mainly because, without appropriate updates, an AI model will decay over time. Thus, defining an update mechanism minimises this problem. Organisations should discuss how and when the model needs to be retrained to be effective. For example, an organisation might decide to retrain manually or apply continuous learning every week. Furthermore, this involves specifying how new data is incorporated into the model.

Define the problem to resolution process

The decay of AI systems surges amid serving errors. Organisations may want to respond by releasing new updates that address known issues and introduce new functionality. Thus, it is key to define non-regular update cycles, including how often this should be done, who is responsible, and how the model is tested and integrated with the existing solution. This can be achieved by adopting a problem resolution process, which consists of keeping a record of the issue, the solution, its rationale, the type of fix (e.g., permanent, or short term), and its effectiveness. This supports having dialectic and challenging discussions with stakeholders in the AI use case. It is worth noting that major functionality updates are better done by conceiving them as a new AI use case and following all the stages starting from Development (see Section 7.3).

7.7 The retirement stage

At some point in time, organisations may decide to take an AI system out of service, either because they want to withdraw active support, or due to partial or total replacement [24]. Thus, this step involves deactivating, disassembling and removing the AI system. Furthermore, when the AI is to be de-operationalised due to replacement, it is necessary to consider how data will be migrated, for example, between companies, and how to transition towards the new system.

Assess deactivation risks

The first stage in deactivating a model is to assess what risks are entailed in taking this step. This relates to both a customer-facing disruption of the service, as well as internal disruptions due to interconnections with other systems. Analogous to the risk assessment prior to launch, a risk assessment should investigate the risks to the users of withdrawing functionality (for example, in safety-critical applications) and/or evaluating risk in transferring sensitive data to other parties.

Handle AI residuals

Once it has been decided how an AI system will be de-operationalised, it is time to consider what to do with its residuals, such as stored data (either for training the model, or resulting predictions), source code and firmware. Thus, this step concerns activities to ensure that AI residuals are handled, replaced, or eradicated appropriately, and to identify critical disposal needs. Practically, this results in identifying what to keep, what and how to discard. IEEE [24] recommends making these decisions based on critical disposal needs to be specified in agreements, policies or resulting from evaluating the environmental, legal, safety and security impact of eradicating or retaining data. For example, an application may require maintaining records for some years as evidence of high-stake AI predictions. Finally, when another system upgrades an AI system, only the impacted AI residuals should be deactivated and removed [24].

8 The rationale for ethics-based auditing of AI systems

8.1 Prevalence and modes of AI ethics failure

To understand how ethics-based auditing (EBA) in general, and **capAI** in particular, can improve the trustworthiness of AI, we first need to understand the modes of AI ethics failure. Unfortunately, a few cases of high-profile failure dominate much of the relevant debate. Comprehensive reviews of cases of failure are lacking so far. However, understanding the nature and extent of AI ethics failure is paramount for any ethics audit protocol to be effective. **capAI** is based on a comprehensive review of prevalent ethical failures of AI systems [15]. We have collected and analysed 106 cases from all across the globe where AI systems have caused public controversy by violating social norms and values. The median date of incidence is 2017, with the earliest case dating back to 2011, showing how recent the AI ethical failure phenomenon is. Our analysis highlights three main modes of failure.

The most common mode of AI ethics failure is privacy intrusion, accounting for half of our cases. Privacy has recently become a much higher preoccupation for stakeholders. Regulatory interventions such as the EU's 2016 General Data Protection Regulation and the 2018 California Consumer Privacy Act have made consumers more aware of their rights to safeguarding data privacy. There are two related failures embedded here: *consent to use the data* and *consent to use the data for the intended purpose*. Privacy violation can also occur when data is obtained with consent, but is then used for a purpose not consented to.

The second most common mode of AI ethics failure is algorithmic bias, accounting for 30% of our cases. It refers to reaching a prediction that systematically disadvantages (or even excludes) one group based on personal identifiers such as race, gender, sexual orientation, age or socio-economic background. Biased AI prediction can become a significant threat to fairness in society, especially when attached to institutional decision-making [89, 90]. While analysing cases of AI bias, we also found some that were more difficult to assess. In particular, we encountered cases where the bias exhibited might align with users' preferences. For instance, many dating apps tend to recommend same-race dates to users as they found that users themselves prefer to date people of their own race. In such cases, while bias is evident, it is perhaps less easy to attribute a failure tag. Bias is indisputable when it comes to unequal opportunities or treatment for a proportion of people, like excluding specific gender or race from working, education and financial opportunities. However, when it involves users' preferences, organisations face a reputationally more contested choice between reinforcing the existing bias in user preferences or choosing to take affirmative action to correct such biased preferences.

The third major mode of AI ethics failure arises from the problem of explainability, otherwise known as 'explainable AI' or 'X-AI'. These account for 14% of our cases. Here, AI is often described as a 'black box' from which people cannot explain the decision that the AI algorithm has reached. AI systems are often described as opaque, as their statistical learning obscures the understanding and assessment of their predictions. Illustrations of algorithmic opacity are easy to conjure: Why did the AI reject this loan application? Why is the AI confident that this patient has cancer? Why is the AI showing this advertising campaign to

this person? Algorithmic opacity raises questions on how to verify the quality of algorithm predictions to ensure users can trust them. For one, if not handled properly, algorithms may reinforce (i) epistemic and (ii) normative concerns that can lead to unfair outcomes [41]. The criticisms stem from the fact that people are usually only informed of the final decisions made by AI, whether that be loan grants, university admission or insurance prices, but at the same time have no idea how or why the decisions are made. The question of explainability arises when humans are affected adversely by the prediction made by an AI system, or in extreme cases, are harmed by AI-based decisions. The case of injuries or deaths resulting from autonomous vehicles is commonly cited.

Looking across all types of AI failure, the most frequent problems are privacy and bias (see Figure 10). Together, they amount to more than four out of five cases of failure. The common theme that runs across these failures is the integrity of the data used by the AI system. AI systems work best when they have access to large datasets. Organisations face significant temptations to acquire and use all the data they have access to, irrespective of users' consent (what is also known as 'data creep') or neglect the fact that customers have not given their explicit consent for this data to be used for a specific purpose (what is also known as 'scope creep'). In both cases, the firm violates the customer's rights to privacy by using data it had not been given consent to use in the first place, or to use for the purpose at hand.

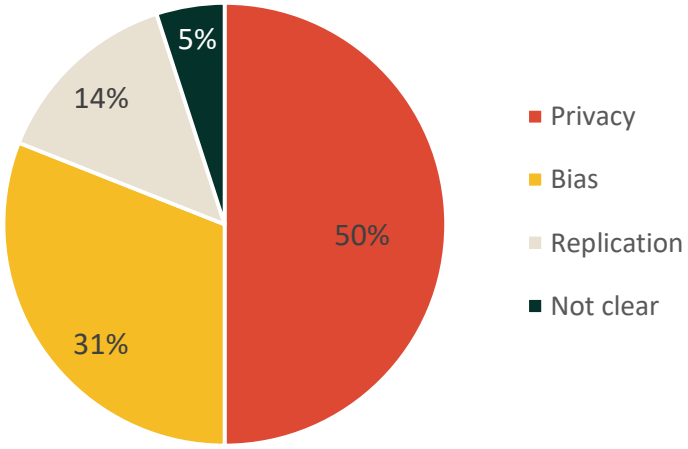

Figure 10: *Incidence of AI failure modes (n=106 case)*

The bias problem is often referred to as 'algorithmic bias' [91] – yet the algorithm, of course, is not at fault here. Algorithms are value-free and inherently agnostic. Grasping the contextual nature of protected variables, such as age, race, gender and sexual orientation, requires a cognitive understanding that is beyond the reach of AI systems. The root cause for algorithmic bias rests firmly with programmers and the veracity and relevance of the data they use. Bias can emerge when customer preferences shift, and machine learning models are not retrained. As they work with increasingly outdated data (which they were trained on), their predictions become biased (a phenomenon also referred to as 'model creep'). However,

even with up-to-date data, AI models can 'learn' from the inherent bias in the real-world data, so that their prediction can reinforce or replicate the existing bias.

In summary, AI ethical failure is considerably more prevalent than commonly assumed. Also, most cases of ethical failure have occurred in the last five years, showing the clear need for new governance mechanisms to ensure that AI systems are legally compliant, technically robust, and adhere to ethical norms and values.

8.2 Conformity assessment and post-market monitoring as stipulated in the AIA

The AIA published by the European Commission on 21 April 2021 is – as mentioned in the introduction – the first attempt to elaborate a general legal framework for AI carried out by any major economy. Importantly, the AIA takes a risk-based approach to AI governance, whereby some AI use cases will be banned entirely. This includes the prohibition of AI systems used for general-purpose social credit scoring and real-time remote biometric identification of a person in public spaces for law enforcement. In contrast, AI systems that pose minimum or no risk (such as spam filters and mobile gaming applications) will not be subject to any obligations under the AIA.

A wide range of so-called 'high-risk' AI systems exists between these two extremes. Technology providers will have to demonstrate that the AI systems they design or deploy adhere to the requirements stipulated in the AIA before placing these systems on the European market. Ultimately, the legal requirements are the same for all high-risk AI systems. According to ANNEX IV in the AIA, these include, among others, obligations on the provider to:

1. document the intended purpose of the AI system in question;
2. provide detailed user instructions;
3. disclose the methods used to develop the system; and
4. justify the critical design choices made by the provider.

The AIA includes several mechanisms designed to ensure that technology providers adhere to the above requirements. Most notably, the providers of high-risk AI systems may face hefty fines if they fail to comply with the requirements stipulated in the AIA. For example, non-compliance with the prohibition of specific uses of AI systems may subject providers to fines of up to €30m, or 6% of their total annual turnover, whichever is higher.

For our purposes, the two most important governance mechanisms referred to in the AIA are *conformity assessments* and *post-market monitoring*. Through conformity assessments, providers can show that their high-risk systems comply with the requirements set out in the AIA ex-ante, i.e., before placing the system on the market. Once a high-risk AI system has demonstrated conformity with the AIA – and received a so-called CE marking – it can be deployed in, and move freely within, the internal EU market. Post-market monitoring refers to

the requirement that providers must document and analyse the performance of high-risk AI systems throughout their lifetimes.

Importantly, the AIA does not specify *how* conformity assessments should be conducted in practice. However, it does give some guidance on *who* has to do conformity assessments and *when*. There are three ways in which these conformity assessments can be conducted. Which type of conformity assessment is appropriate depends on the nature of the high-risk AI system.

Consider the many high-risk AI systems used as safety components of consumer products that are already subject to third-party, ex-ante conformity assessments under current product safety law. These include, for example, AI systems that are part of medical devices or toys. In these cases, the requirements set out in the AIA will be integrated into existing sectoral safety legislation. This avoids duplicating administrative burdens and maintains clear roles and responsibilities while ensuring a strong consistency among the different strands of EU legislation. However, it also implies that no 'AI specific' conformity assessments will occur. Instead, compliance with the AIA will be assessed through the third-party conformity assessment procedures already established in each sector.

High-risk AI systems that do not fall into the first category are called 'stand-alone' systems. The complete list of stand-alone, high-risk AI systems subject to conformity assessments is found in ANNEX III to the AIA. These include AI systems used in recruitment, determining access to educational institutions, and profiling persons for law enforcement. Providers of stand-alone, high-risk AI systems have two options for conducting ex-ante conformity assessments. They can either (a) conduct ex-ante conformity assessments based on internal control or (b) involve a third-party auditor (i.e., a notified body) to assess their quality management system and technical documentation.

It should be noted that procedure (a) is only an option where the AI system is fully compliant with the requirements set out in Chapter 2 of Title III of the AIA. When, in contrast, the compliance is only partial – or harmonised standards do not yet exist – providers are obliged to follow procedure (b). This may seem opaque. However, Figure 11 (below) illustrates through a simple flow-chart different ways for conducting conformity assessments.

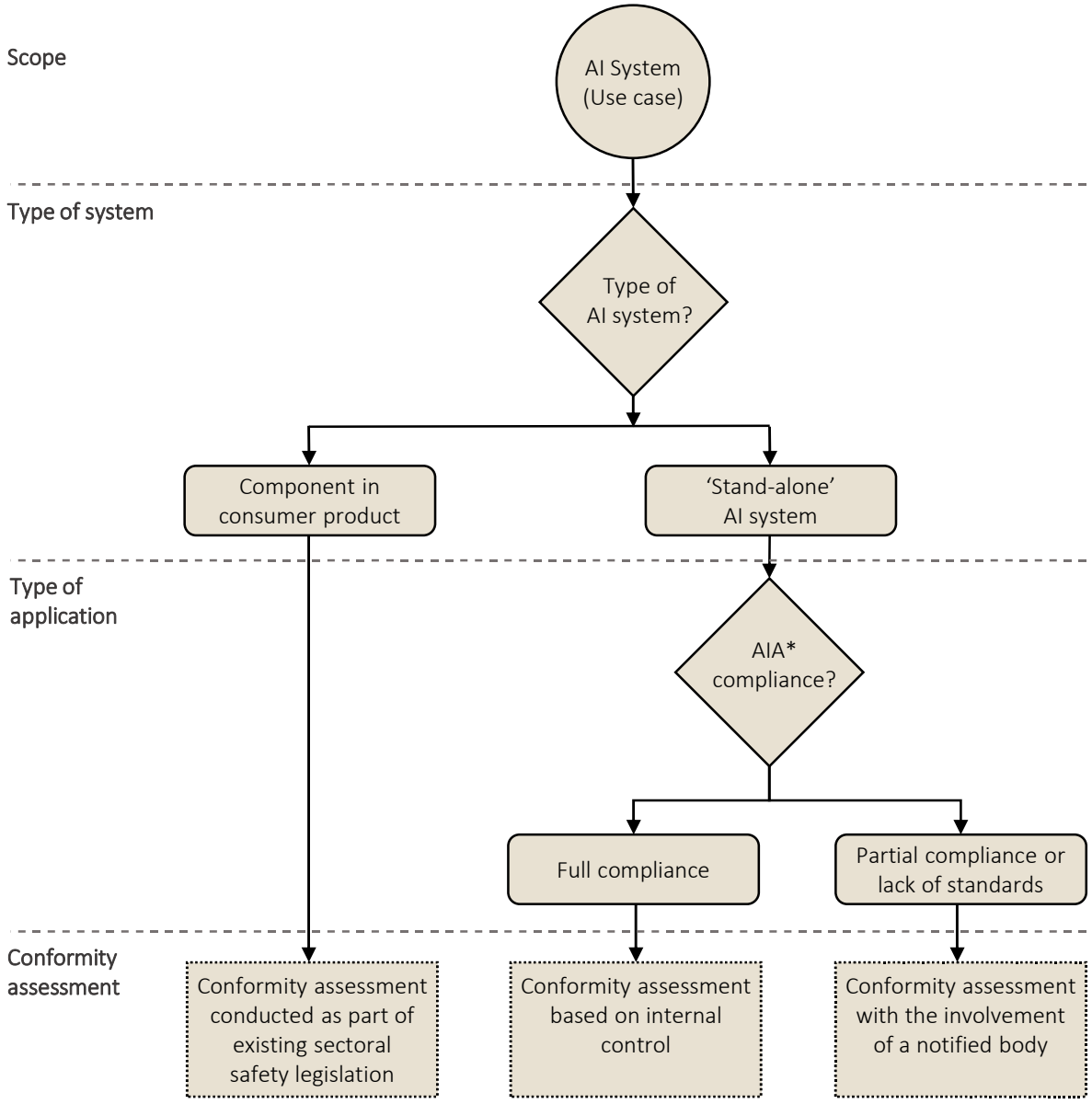

Figure 11: *Ways to conduct conformity assessments for high-risk AI systems [14]*

In addition to the ex-ante conformity assessments described above, providers of high-risk AI systems are also expected to establish and document post-market monitoring systems. The task of post-market monitoring is to document and analyse the behaviour and performance of high-risk AI systems throughout their lifetime. These ex-post assessments are crucially complementary to ex-ante certifications, since providers of high-risk AI systems are expected to report any serious incidents or any malfunctioning that constitute a breach of Union law. They are also obliged to take immediate and corrective actions needed to bring the AI system under conformity or withdraw it from the market.

To detect, report on and address system failures in effective and systematic ways, providers must first draft post-market monitoring plans that account for, and are proportionate to, the nature of their respective AI systems. The post-market monitoring plan is, in turn, part

of the required documentation that constitutes the basis for the conformity declaration. Here, it is important to note that such ongoing, post-market monitoring is intrinsically linked to quality management as a whole. According to the AIA, the main objective of the quality management system is to establish procedures for how high-risk AI systems are designed, tested and verified. However, it should also include procedures for data management and record keeping, and procedures for implementing and maintaining post-market monitoring of the high-risk AI system in question.

Legally mandated, post-market monitoring adds a new element and new complexities to corporate quality management systems. Since providers of high-risk AI systems are not necessarily the ones using them, they must give users clear instructions on the operation of high-risk AI systems, and cooperate with users to enable effective post-market monitoring. For example, providers can consider the requirement that high-risk AI systems shall be designed with capabilities to record automatically (or 'log') their operations and decisions. As per contractual agreements, these logs can be controlled by the user, the provider or a third party. In any case, however, it is the provider's responsibility to ensure *that,* and plan for *how,* high-risk AI systems automatically generate logs.

Combined, the ex-ante conformity assessments and the post-market monitoring mandated by the AIA constitute a coordinated and robust approach basis for enforcing the proposed EU regulation. However, the AIA only contains limited guidance on how to conduct conformity assessments and post-market monitoring in practice, and an enforcement mechanism will only be as good as the institution backing it. Thus, we next turn to examine the institutional structure proposed in the AIA.

8.3 Roles and responsibilities in an emerging European auditing ecosystem

Ensuring that high-risk AI systems satisfy the various requirements set out in the AIA would require a well-developed auditing ecosystem consisting of two components. First, an institutional structure is needed that clarifies the roles and responsibilities of private companies, national and supranational authorities. This would also include ensuring accountability for different types of system failures. Second, the actors in the ecosystem need access to well-calibrated auditing tools and the necessary expertise to carry out the process and show that high-risk AI systems comply with the AIA. Such an ecosystem does not yet exist. Nevertheless, the AIA sketches the contours of an emerging European AI auditing ecosystem.

According to the AIA, the providers and users of high-risk AI systems share the responsibility for ensuring compliance and identifying and mitigating potential breaches of compliance. However, to ensure regulatory oversight, the Commission proposes to set up a governance structure that spans both Union and national levels. A 'European Artificial Intelligence Board' will be established at a Union level to collect and share best practices among member states and issue recommendations on uniform administrative practices. In

addition, the Commission will set up and manage a centralised database for registering stand-alone, high-risk AI systems. The purpose of the database is to increase public transparency and enable ex-post supervision by competent authorities.

At a national level, member states will have to designate a competent national authority to supervise the application and implementation of the AIA. Importantly, this national supervisory authority should not conduct any conformity assessments itself. Instead, it will act as a notifying authority that assesses, designates and notifies third-party organisations that, in turn, conduct conformity assessments of providers of high-risk AI systems. In the proposed EU legislation, these third-party organisations are sometimes referred to as 'conformity assessment bodies', but, in other contexts, they are often simply called 'notified bodies'. To become a notified body, an organisation must apply to the notifying authority of the member state in which they are established.

The main task of a notified body is to assess and approve the quality management systems that providers of high-risk AI systems use in the process of design, development and testing. Further, the notified body shall examine the technical documentation for each high-risk AI system produced under the same quality management system. Based on these assessments, the notified body shall then determine whether the quality management system and the technical documentation satisfy the requirements set out in the AIA. The notified body shall issue an EU technical documentation assessment certificate where conformity has been established. Figure 12 below provides an overview of the relationship between different private organisations and institutional bodies in the process of assessing and certifying stand-alone, high-risk AI systems.

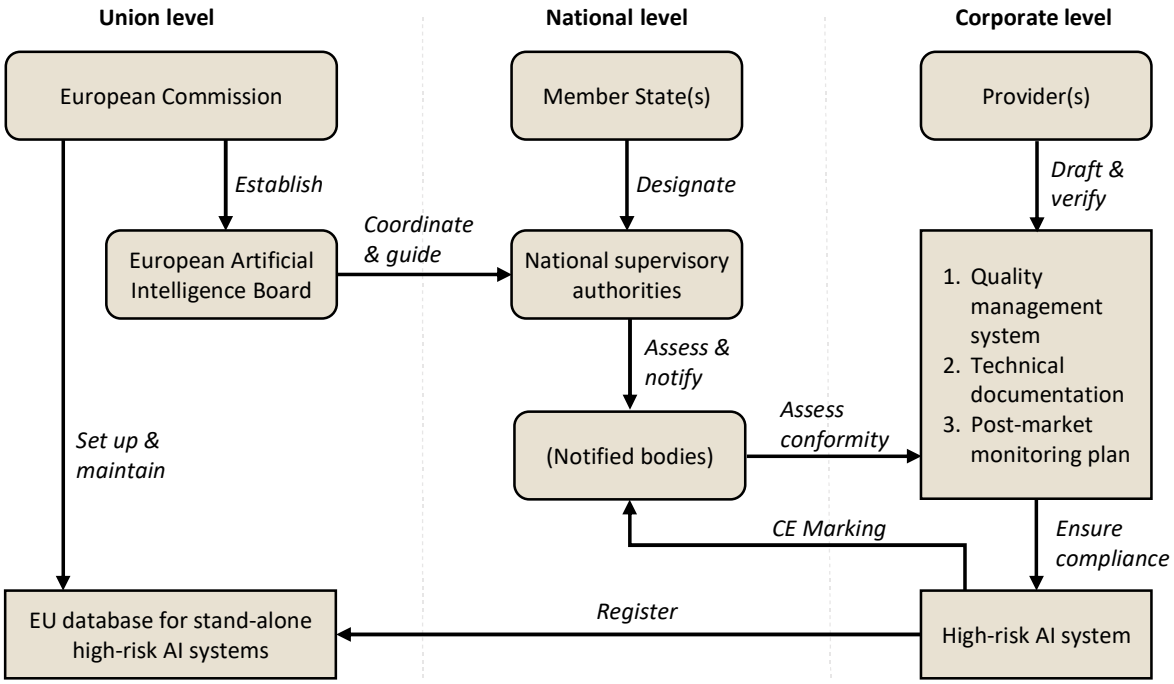

Figure 12: Roles and responsibilities during conformity assessments with the involvement of third-party auditors

Admittedly, Figure 12 gives a somewhat idealised picture of the roles and responsibilities outlined in the AIA. First of all, the relationships – here indicated by directional arrows – are in reality bidirectional. For example, although the national supervisory authority is responsible for assessing and notifying conformity assessment bodies, it does so based on the application and material submitted by organisations that wish to be notified. Similarly, while the notified body is responsible for conducting conformity assessments, high-risk AI systems providers must give the notified body timely access to all resources and documents that are necessary for a comprehensive assessment to take place, and report any severe incidents or malfunctioning of their high-risk AI systems directly to the national surveillance authority. To deliver on these expectations, providers and users of AI systems will have to update their internal quality management systems and appoint new roles within their organisations.

8.4 The remaining gap

While the logic behind the conformity assessments and the post-market monitoring activities mandated in the AIA is clear, many details concerning how these should be conducted in practice have yet to be spelt out. Moreover, the AIA focuses exclusively on AI systems aimed at the market. Taken together, AI technology providers need further procedural guidance on how they can verify claims made about the AI systems they design and deploy. This is where **capAI** comes in. As mentioned in the introduction, **capAI** has been developed with two use cases in mind. First, providers of 'high-risk' AI systems may use **capAI** to show compliance with the EU's Artificial Intelligence Act (AIA). Second, providers of 'low-risk' AI systems, i.e., systems that do not fall within the regulatory scope of the AIA, may use **capAI** to operationalise their commitments to voluntary codes of conduct. To serve these functions, **capAI** draws on EBA, a governance mechanism already established in the academic literature. The following section expands on what EBA is, and how it works.

9 The implementation of ethics-based auditing

9.1 Introduction

The theoretical foundation and procedural blueprint for **capAI** is ethics-based auditing (EBA). At a high level of abstraction, EBA is a governance mechanism that allows organisations to operationalise their ethical commitments and validate claims made about their AI systems [13]. Due to these affordances, EBA can help organisations show compliance with the requirements set out in the AIA and ensure that the AI systems they design and deploy adhere to voluntary codes of conduct. Operationally, EBA is characterised by a structured process whereby an entity's present or past behaviour is assessed for consistency with relevant principles or norms. Thus, EBA differs from merely publishing a code of conduct because its main activity consists in demonstrating adherence to a predefined baseline.

EBA has attracted much attention in recent years. For example, regulators like the UK Information Commissioner's Office (ICO) have provided guidance on how to audit AI systems [92]. Moreover, professional services firms, including PwC and Deloitte, technology-based start-ups like ORCAA, and NGOs like ForHumanity, are all developing auditing tools to help clients verify claims about the trustworthiness of their AI systems [93–95]. However, central practical questions have remained unanswered despite a growing interest in EBA from policymakers and organisations. These questions include: Who should conduct the audits? According to which metrics should AI systems be evaluated? And, how can EBA be integrated into existing organisational governance structures? The remainder of this section describes how **capAI** helps fill these critical knowledge gaps by drawing on previous research. However, first, the key terms must be defined.

9.2 Defining key terms

We use the term *governance mechanism* to demarcate the set of activities, structures and controls wielded by various parties to exert influence and achieve normative ends. With governance, we understand a process whereby elements in society wield power, authority and influence, and enact policies. Governance thus consists of both hard and soft aspects. Hard governance mechanisms are systems of rules elaborated and enforced through institutions to govern agents' behaviour. Soft governance embodies mechanisms that exhibit some degree of contextual flexibility, like subsidies and taxes.

A distinction can also be made between formal and informal governance mechanisms. Formal governance mechanisms are officially stated, communicated and enforced. While hard governance mechanisms are formal by definition, not all formal governance mechanisms are necessarily hard. Budgets, codes of conduct and reward criteria are, for example, soft yet formal governance mechanisms. Informal governance comprises common values, beliefs and traditions that direct the behaviour of individuals and groups within organisations. However, the latter is particularly relevant because decisions made by AI systems may be deserving of scrutiny even when they are not illegal.

With *EBA*, we refer to a soft yet formal governance mechanism that can be used by various parties to control or influence the behaviour of organisations and systems. As mentioned in the introduction, EBA is characterised by a structured process whereby an entity's present or past behaviour is assessed for consistency with relevant principles or norms. Throughout this purpose-oriented process, various *tools* (such as software programs and standardised reporting formats) and *methods* (like stakeholder consultation or adversarial testing) are employed to verify claims and create traceable documentation. Naturally, different EBA processes employ different tools and contain different steps. The protocols that govern specific EBA processes are hereafter referred to as auditing *frameworks*. This use of the terms *auditing*, *tools* and *frameworks* is in keeping with the Institute of International Auditors [96].

9.3 Background

The idea of auditing software is not new. Since the 1970s, computer scientists have been involved in research addressing issues of certifying software according to functionality and reliability [97]. Nor is the idea of auditing AI systems for consistency with societal norms new. Since popularised by Sandvig and colleagues [98, 99], EBA has attracted much attention from policymakers and academic researchers alike. However, before returning to the more recent literature, something should be said about where the idea came from.

As a governance mechanism, auditing has a long history of promoting trust and transparency in security and financial accounting. Valuable lessons can be learned from these domains. One is that the process of auditing is always purpose-oriented. For EBA, the purpose is to ensure that AI systems operate in ways that align with specific guidelines (such as the requirements on high-risk AI systems stipulated in the AIA, in the case of **capAI**). Another lesson is that auditing presupposes operational independence between the auditor and the auditee. Whether the auditor is a government body or a third-party contractor, the main point is to ensure that the audit is run independently from the regular chain of command within organisations. The reason for this is to minimise the risk of collusion between auditors and auditees and to allocate responsibility for different types of harm or system failures.

To conceptualise the roles and responsibilities of different stakeholders throughout the process of EBA, we build on a framework (see Figure 13 below) developed by the Institute of Internal Auditors (IIA). According to this framework, the principal stakeholders include *organisations* that design and deploy AI systems (who are accountable for the behaviour of their systems), the *management* of such organisations (who are responsible for achieving organisational goals, including adhering to ethical values), independent *auditors* (who are tasked with objectively reviewing and assessing how well an organisation adheres to relevant principles and norms), and *regulators* (who are monitoring the compliance of organisations). Note that this framework is akin to – and compatible with – the emerging AI auditing ecosystem sketched by the European Commission in the AIA, which is displayed in Figure 13 below.

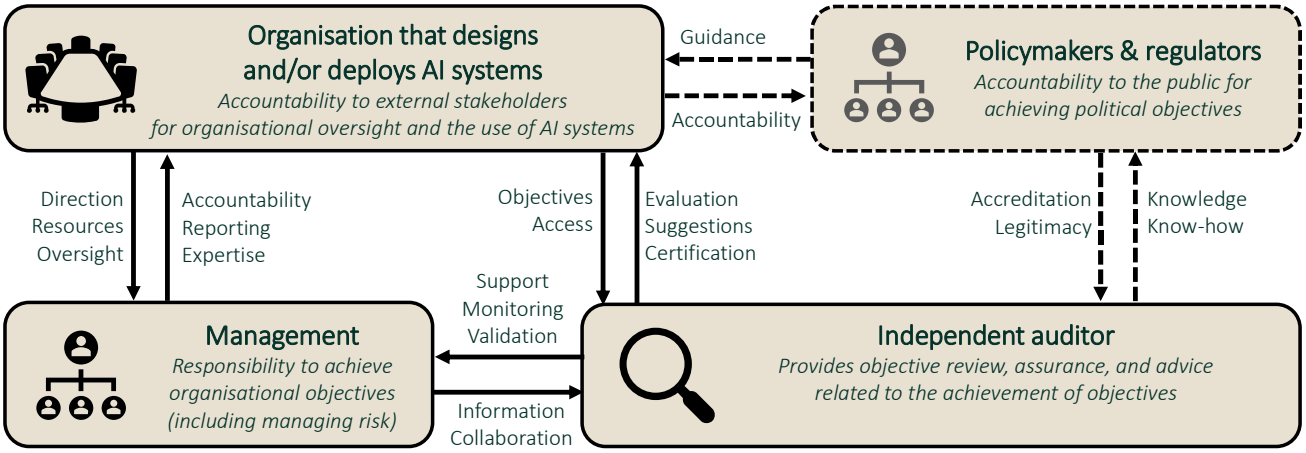

Figure 13: Conceptual sketch of the roles and responsibilities during independent audits. Source: IIA, 2017

Despite the similarities with audits in other sectors, it should be recognised that widely accepted standards for EBA of AI systems have yet to emerge. Nevertheless, it is possible to distinguish some emerging approaches to EBA. *Functionality audits*, for example, focus on the rationale behind decisions made by AI systems. In contrast, *code audits* entail reviewing the source code of an AI system. Finally, *impact audits* investigate the types, severity and prevalence of effects of an AI system's outputs. The latter is particularly relevant: since autonomous and self-learning AI systems may evolve and adapt over time as they interact with their environments, EBA needs to include (at least elements of) continuous, real-time monitoring.

All the above approaches are complementary and can be combined in designing EBA frameworks. Auditing *frameworks* are protocols that describe a specific auditing procedure, define what is to be audited, by whom, and according to which standards. Typically, auditing frameworks originate from one of two processes. The first type consists of national or regional strategies. At a EU level, several documents have been published that are relevant for the EBA of AI systems. For example, the AI HLEG published an *Assessment List for Trustworthy AI* that organisations may incorporate into existing governance structures [100].

The second type of auditing frameworks originates from the expansion of data regulation authorities to account for the effects that AI systems have on informational privacy (see, e.g., the CNIL's *Privacy Impact Assessment,* [101]). The experience data regulation agencies have of translating principles into governance protocols provides valuable blueprints for auditing AI systems. However, auditing frameworks with roots in data regulation tend to account only for specific ethical concerns, such as those related to privacy. While safeguarding personal data, The *General Data Protection Regulation* (GDPR), for example, does not offer protection from inaccurate or unfair outcomes of AI systems. This calls for caution since an exclusive focus on one, or even a few, ethical challenges risks leading to sub-optimisation on a system level.

To synthesise, auditing frameworks converge around a procedure based on impact assessments. IAF [102] summarised this procedure in eight steps. These steps constitute the procedural blueprint not only for EBA in general, but also for **capAI** in particular:

1. Describe the purpose of the AI systems.
2. Define the standards or verifiable criteria based on which the AI systems should be assessed.
3. Disclose the process, including a full account of the data, data use and parties involved.
4. Assess the impact the AI systems have on humans and the environment.
5. Evaluate whether the benefits and mitigated risks justify the use of AI systems.
6. Determine the extent to which the system is reliable and transparent.
7. Document the results and considerations.
8. Reflect and evaluate, i.e., create a feedback loop.

In contrast to procedural frameworks, auditing *tools* are conceptual models or software products that help measure, evaluate or visualise one or more properties of AI systems. The IRP and the ESC provided by **capAI** are thus examples of tools designed to enable and facilitate EBA of AI systems. However, a great variety of such tools have already been developed by both academic researchers and privately employed data scientists, and reviewing these holds valuable lessons.

To start with, different tools help ensure the ethical alignment of AI systems in different ways. Some tools facilitate the audit process by visualising the output from AI systems. FairVis, for example, is a visual analytics system that integrates a subgroup discovery technique, thereby informing normative discussions about group fairness [103]. Another example is Fairlearn, an open-source toolkit that treats any AI system as a black box. Fairlearn's interactive visualisation dashboard helps users compare the performance of different models [104]. The main takeaway here is that visualisation helps developers and auditors to create more equitable algorithmic systems. Other tools improve the interpretability of complex AI systems by generating more straightforward rules that explain their predictions. For example, Shapley Additive exPlanations (SHAP) calculates the marginal contribution of the features underlying a model's prediction [105]. The explanations provided by tools like SHAP are useful, for example when determining whether protected features have unjustifiably contributed to a decision made by AI systems.

Auditing tools have also been developed to help democratise the study of AI systems. Consider the TuringBox, which was developed as part of a time-limited research project at MIT. This platform allowed software developers to upload the source code of an AI system, to let others examine it [106]. The TuringBox provided an opportunity for developers to benchmark their system's performance regarding different properties. Other auditing tools

help organisations document the software development process and monitor AI systems throughout their life cycle. For example, AI Fairness 360 developed by IBM includes metrics and algorithms to monitor, detect and mitigate bias in datasets and models [107]. Finally, some tools have been developed to aid developers in making pro-ethical design choices by providing useful information about the properties and limitations of AI systems. Such tools include end-user licence agreements and datasheets [108].

In short, a wide variety of auditing frameworks and tools have already been developed to help organisations and societies manage the ethical risks posed by AI systems. However, these tools are often employed in isolation. Hence, to be feasible and effective, auditing procedures need to combine existing conceptual frameworks and software tools into a structured process that monitors each stage of the software development life cycle to identify and correct the points at which ethical failures (may) occur. Therefore, **capAI** combines elements of functionality-, code- and impact auditing. This does not constitute a break with the periodic nature of traditional audits but rather a methodological evolution.

Taken together, previous work holds important lessons. Building on experience from financial audits, **capAI** takes as a starting point that the primary responsibility for identifying and executing steps to ensure that AI systems are ethically sound rests with the management of the organisations that design and operate such systems. In contrast, the independent auditor's responsibility is to assess and verify claims made by the auditee about its processes and AI systems, and to ensure that there is sufficient documentation to respond to potential enquiries from public authorities or decision-making subjects. Moreover, building on best practices from quality management in software development, a critical function of the IRP is to spark and inform ethical deliberation throughout the software development process. The main idea here is that continuous monitoring and assessment ensure that a constant flow of feedback concerning the ethical behaviour of AI systems is worked into the next iteration of its design and application.

9.4 How capAI harnesses the promise of ethics-based auditing

As mentioned above, **capAI** takes EBA as a procedural blueprint. The reason for this is that EBA offers several methodological affordances.

First, EBA can help relieve human suffering by anticipating potential negative consequences before they occur. To establish safeguards against unexpected, unwanted or unknown behaviours, EBA should combine minimum requirements on system performance with automated control of an AI system's output. This is why **capAI** emphasises that technology providers conduct and document ethical impact assessments already in the concept stage of the software development process.

Second, EBA can improve user satisfaction and unlock economic growth by building trust in available technologies through procedural transparency, documentation and actionable explanations. Even when algorithms are opaque, AI systems can be understood intentionally, through their design and in terms of their inputs and outputs. This logic is reflected by the fact

that the IRP demands technology providers should define the purpose of the intended use case and key performance indicators upfront.

Third, EBA can help ensure accountability by tapping into existing internal and external governance structures. EBA also helps clarify the roles and responsibilities of different stakeholders beyond the auditor, including those system owners and existing civil institutions, so that responsibility for different types of system failures can be allocated. For this reason, the IRP helps organisations define the roles and responsibilities of top managers, product owners, project managers and data scientists in relation not only to the software development process but also to external stakeholders.

Fourth, EBA facilitates local alignment of ethics and legislation. Laws and regulations differ between geographies and operational sectors. EBA can account for different types of AI-related ethical harms by identifying and communicating errors, tensions and risks while adopting sector-specific standards. In short, while **capAI** can be used by an organisation to demonstrate adherence to the requirements on high-risk AI systems stipulated in the AIA, the same procedure can also be leveraged to operationalise commitments to other organisational values, including voluntarily adopted codes-of-ethics.

Fifth, EBA can provide decision-making support to executives and legislators by defining and monitoring outcomes. The process of defining goals and evaluation criteria for audits forces AI practitioners to consider upfront the normative ends of the systems they develop. Moreover, EBA can help understand which normative values are embedded in a system. For example, the ESC provided by **capAI** summarises relevant information about the AI system and makes this information easily available to users, consumers and citizens to act upon.

Sixth, EBA can help balance conflicts of interest. For example, EBA can provide a basis for accountability while preserving the integrity of intellectual property rights, e.g., by containing access to sensitive information to authorised third-party auditors. This is why **capAI** consists of an IRP and an ESC. It is important to stress that the reason for introducing different layers of transparency is not to create opacity. Rather, it is to create a trusted environment in which organisational learning and continuous improvements in the design of AI systems can take place.

These benefits are all potential, not guaranteed, and depend on how EBA is operationalised and external environmental factors. The key success factor is how EBA procedures are designed and implemented in practice. For this reason, we now turn to outline the best practices on which **capAI** is based.

9.5 Best practices for successful implementation

EBA procedures need not be difficult to implement. However, EBA procedures must be informed by existing best practices to be feasible and effective in practice. Some of these best practices are abstract and relate to how stakeholders view EBA. Others are tangible and concern the specific design of individual EBA procedures. As a starting point, it should be acknowledged that AI systems are not isolated technologies. Rather, AI systems both help

shape and are shaped by larger sociotechnical systems. Hence, system output cannot be considered biased or erroneous without some knowledge of the available alternatives. Therefore, holistic approaches to EBA must seek input from diverse stakeholders, e.g., for an inclusive discourse about key performance indicators (KPI). However, regardless of which KPI an organisation chooses to adopt, audits are only meaningful insofar as they enable organisations to verify claims made about their AI systems. This implies that EBA procedures themselves must be traceable. By providing a traceable log of the steps taken in designing and developing AI systems, audit trails can help organisations verify claims about their engineered systems.

Further, to ensure that AI systems are ethically sound, organisational policies need to be broken down into tasks for which individual agents can be held accountable. By formalising the software development process and revealing (parts of) the causal chain behind decisions made by AI systems, EBA helps clarify the roles and responsibilities of different stakeholders, including executives, process owners and data scientists. However, allocating responsibilities is not enough. Sustaining a culture of trust also requires that people who breach ethical and social norms are subject to proportional sanctions. At the same time, doing the right thing should be made easy. This can be achieved through strategic governance structures that align profit with purpose. The 'trustworthiness' of a specific AI system is never just a question about technology but also about value alignment. In practice, this means that the checks and balances developed to ensure safe and benevolent AI systems must be incorporated into organisational strategies, policies and reward structures.

Importantly, EBA does not provide an answer sheet but a playbook. This means that EBA should be viewed as a dialectic process wherein the auditor ensures that the right questions are asked and answered adequately. This means that auditors and system owners should work together to develop context-specific methods. To manage the risk that independent auditors would be too easy on their clients, licences should be revoked from both auditors and system owners in cases where AI systems fail. However, it is difficult to ensure that an AI system contains no bias, or to guarantee its fairness. The goal from an EBA perspective should, therefore, be to provide useful information about when an AI system is causing harm or when it is behaving in a way that is different from what is expected. This pragmatic insight implies that audits need to monitor and evaluate system outputs continuously, i.e., through 'oversight programmes', and document performance characteristics in a comprehensible way.

Finally, the alignment between AI systems and specific ethical values is a design question. Ideally, properties like interpretability and robustness should be built into systems from the start, e.g., through 'Value-Aligned Design'. However, the context-dependent behaviour of AI systems makes it difficult to anticipate the impact AI systems will have on the complex environments in which they operate. By incorporating an active feedback element into the software development process, EBA can help inform the continuous re-design of AI systems. Although this may seem radical, it is already happening: most sciences, including engineering and jurisprudence, not only study their systems, but they also simultaneously build and modify them.

Taken together, these generalisable lessons suggest that EBA procedures – even when imperfectly implemented – can make a real difference as to how AI systems are designed and deployed. Building on the best practices discussed in this section, **capAI** has been developed to meet the following criteria:

1. **Holistic**, i.e., treat AI systems as an integrated component of larger sociotechnical contexts.
2. **Traceable**, i.e., assign responsibilities and document decisions to enable follow-up.
3. **Accountable**, i.e., help link unethical behaviours to proportional sanctions.
4. **Strategic**, i.e., align ethical values with policies, organisational strategies and incentives.
5. **Dialectic**, i.e., view EBA as a constructive and collaborative process.
6. **Continuous**, i.e., identify, monitor, evaluate and communicate system impacts over time.
7. **Driving re-design**, i.e., provide feedback and inform the continuous re-design of AI systems.

Of course, these criteria are aspirational and, in practice, unlikely to be satisfied all at once. Nevertheless, we must not let perfect be the enemy of good. Organisations are thus advised to work in the spirit of the above criteria when employing **capAI**.

9.6 Managing known limitations and pitfalls

The extent to which **capAI** can contribute to ensuring that AI systems behave ethically depends not only on how auditing procedures are designed but also on the intent of different stakeholders. An analogy, borrowed from Floridi [109], is helpful to illustrate this point: the best pipes may improve the flow but do not improve the quality of the water, yet water of the highest quality is wasted if the pipes are rusty or leaky. As the pipes in the analogy, no EBA procedure is morally good in itself. However, they can realise moral goodness if adequately designed and combined with the right values.

That said, even the best efforts to translate ethical principles into organisational practices may be undermined by a set of ethical risks. Floridi [110] lists five such risks. For our purposes, the three most relevant of these risks are: *ethics shopping*, i.e., the malpractice of cherry-picking ethics principles to justify pre-existing behaviours; *ethics bluewashing*, i.e., the malpractice of making unsubstantiated claims about the ethical behaviour of an organisation or an AI system; and *ethics lobbying*, i.e., the malpractice of exploiting 'self-governance' to delay or avoid necessary legislation about the design of AI systems.

Of course, EBA procedures (such as **capAI**) are not immune to these concerns. For instance, consider the keen interest taken by large technology companies in developing tools and methods for EBA. While commendable, experiences from self-governance initiatives in

other sectors suggest that the industry may not want to reveal insider knowledge to regulators, but instead use its informational advantage to obtain weaker standards. However, the fact that EBA does not resolve all tensions associated with the governance of AI systems is not necessarily a failure. Rather, it is to the credit of EBA that it helps manage some of these tensions. For example, by demanding that ethical principles and codes of conduct are clearly stated and publicly communicated, EBA ensures that organisational practices are subject to additional scrutiny, which, in turn, may counteract ethics shopping. Similarly, when conducted by an independent auditor and provided that the results are publicly communicated, EBA can also help reduce the risk of ethics bluewashing by allowing organisations to validate the claims made about their ethical conduct and the AI systems they operate.

The challenges related to ethics shopping and ethics bluewashing discussed above apply to all attempts to implement AI governance in practice. However, there is also a range of conceptual, technical, economic and institutional constraints associated with EBA more specifically. Three of them are worth highlighting here.

First, EBA is constrained by the difficulty of quantifying externalities that occur due to indirect causal chains over time. For example, while practitioners are encouraged to consider – and account for – the social implications of a prospective AI system throughout the software development process, this is often difficult in practice since scalable and autonomous systems may have indirect impacts that spill over borders and generations. Hence, rather than attempting to codify ethics, one function of auditing is to arrive at resolutions that, even when imperfect, are at least publicly defensible.

Second, a weakness of traditional auditing methodologies is the assumption that test environments sufficiently mimic the later application to allow quality assurance. Put differently, there is a tension between the stochastic nature of AI systems and the linear, deterministic nature of conventional auditing procedures. As a result, the same agile qualities that help software developers meet rapidly changing customer requirements also make it challenging to ensure compliance with pre-specified requirements. This implies that EBA must monitor and evaluate performance-based criteria and process-based criteria.

Third, from a social perspective, there is always the potential for adversarial behaviour during audits. The organisation or AI systems that are being audited may, for example, attempt to trick the auditor by withholding information or temporarily adjusting their behaviour. While many auditing frameworks may anticipate adversarial behaviour, so-called ‘management fraud’ can still evade auditors. Similarly, even when audits reveal flaws within AI systems, power asymmetries may prevent corrective steps from being taken.

A final set of limitations – or pitfalls – stem from implementing new governance mechanisms. While good governance is about balancing conflicting interests, it can take time for socially good equilibria to form. Hence, new governance mechanisms often suffer from pitfalls like *tunnel vision*, whereby overregulation may do more harm than good; *random agenda selection*, whereby special interest groups set priorities; and *inconsistency*, whereby

different standards are used to evaluate different options. EBA may suffer from these limitations as a soft yet formal governance mechanism.

First, let us consider the risk of tunnel vision. It is true that decisions based on incomplete or biased data may end up being erroneous or discriminatory, and that the abilities of AI systems to draw non-intuitive inferences may infringe privacy rights. Nevertheless, when governing new technologies like AI systems, we must be careful not to optimise a single value at the expense of others. Here, the use of multiple evaluation metrics and tolerance intervals can help improve the comprehensiveness of the ethical evaluation, thereby minimising the risk of tunnel vision.

Second, normative values often conflict and require trade-offs. For example, AI systems may improve the overall accuracy but discriminate against specific subgroups in the population. Similarly, different definitions of fairness – like individual fairness, demographic parity and equality of opportunity – are mutually exclusive. Because fundamental political disagreements remain hidden in normative concepts, the development of EBA methodologies runs the risk of random agenda selection, whereby EBA procedures are designed with specific, yet partial or unjustified, normative visions.

Finally, it would be unrealistic to expect decisions made by AI systems to be any less complicated to evaluate from an ethical perspective than those made by humans. Ethical decision-making inevitably requires a frame of reference, i.e., a baseline against which normative judgements can be made. If analysed in a vacuum, AI systems risk being held to higher standards than available alternatives. Such inconsistencies may, in some cases, end up doing more harm than good, as when, for instance, a particular AI system is not used due to concern about accuracy or bias – even if it performs better than humans on the very same measures.

Such absolutism ties back to the naïve belief that we have to – or indeed even can – resolve disagreements in moral philosophy before we start to design and deploy AI systems. A more nuanced approach would be to understand AI systems in their specific contexts and compare them with human decision-makers' relative strengths and limitations. This is also the view that underpins **capAI**.

10 Concluding remarks

AI is a general-purpose technology that is developing rapidly, and its application is increasingly becoming ubiquitous. Thus, ensuring that AI systems behave in an ethical way is of paramount importance. This responsibility lies with the organisations that develop and operate them. In this context, **capAI** should be viewed as a resource – an additional governance mechanism in the management toolbox – that organisations can employ to ensure and show that their AI systems adhere to specific ethical principles. By adopting a process view, **capAI** seeks to promote good software development practices and prevent the most common ethical failures that our research has identified. Regulatory mandates, such as under the AIA, may be seen as an administrative burden. However, the cost of failure in terms of reputational damage, possible legal costs and penalties for non-compliance will, in most cases, outweigh the effort needed to complete the IRP in the first place. In short, **capAI** affords good governance. Following standardised procedures, like IRP, provides organisations with a competitive advantage by pre-empting common failures, validating public claims about ethical AI procedures, and protecting the organisation's reputation in the marketplace.

Further, policymakers are advised to consider EBA as an integral component of multifaceted approaches to managing the ethical risks posed by AI to society. This does not imply that traditional governance mechanisms are superseded. On the contrary, by contributing to procedural regularity and transparency, EBA of AI complements and enhances existing governance mechanisms, like human oversight, certification and regulation. However, this also implies that even in contexts where EBA is necessary to ensure the ethical alignment of AI systems, it is by no means sufficient. For example, it remains unfeasible to anticipate all long-term and indirect consequences of a particular decision made by an AI system. Moreover, while EBA procedures – like **capAI** – can help organisations ensure that their AI system adheres to specific ethics guidelines, how to prioritise between irreconcilable normative values remains fundamentally a political question. Also, how best to implement EBA of AI will vary across different regions and applications. Therefore, a plurality of actors promoting a diverse range of EBA frameworks is needed. Rather than centralising governance, official bodies should retain supreme sanctioning power by authorising independent agencies to, in turn, conduct EBA of AI systems.

capAI codifies current best practices in designing, developing and operating AI systems. However, AI is developing at a fast pace, and new aspects will likely need to be added to the EBA procedures, while some other aspects may no longer be relevant. In the same way as we demand that organisations review and update their AI systems, we acknowledge our responsibility to review and update **capAI**.

Glossary of key terms

AIA: Artificial Intelligence Act. A comprehensive legal framework proposed by the European Commission to govern AI systems within the common market.

Algorithmic bias, see *bias*

ALTAI: The Assessment List for Trustworthy Artificial Intelligence. A practical tool issued by the European Commission to help businesses and organisations self-assess their AI systems' trustworthiness under development**.**

Artificial Intelligence (AI): A non-human program or model that can solve sophisticated tasks.

Artificial Neural Networks (ANN): A type of model for AI inspired by the neural network configurations of the human brain.

Bias: Stereotyping, prejudice or favouritism towards some things, people or groups over others. It can be led by a systematic error in a sampling or reporting procedure, or prejudiced hypotheses made when designing AI models.

Classification: The process of distinguishing between two or more discrete classes already labelled by humans.

Clustering: The process of grouping related examples without existing labels.

Concept drift: The case where the statistical properties of the target variable, which the model is trying to predict, change over time in unforeseen ways. This causes problems because the predictions become less accurate as time passes.

Data creep: The case where AI models seek to incorporate more data and/or different data sources to improve the model's predictive power.

Deep learning (DL): A subset of machine learning, which is essentially a neural network with three or more layers – the input and output layer, and at least one hidden level in between. Modern DL models will have thousands or even millions of hidden layers.

Ethics-based Auditing (EBA): EBA is a governance mechanism that allows organisations to operationalise their ethical commitments and validate claims made about their AI systems.

Expert system: A system that uses AI technology to simulate the judgement and behaviour of a human or an organisation that has expert knowledge and experience in a particular field. Expert systems are generally rule based or deterministic.

Explainability: A set of processes and methods that enables human users to comprehend and trust the results and output created by machine learning algorithms.

GDPR: The General Data Protection Regulation. A European Union law on data protection and privacy.

HLEG: The High Level Expert Group on AI. A group of experts appointed by the European Commission to provide advice on its artificial intelligence strategy.

Hyperparameter: A parameter set to control the learning process, established by the model designer and not learned by the model from data. These parameters can directly affect how well a model trains.

Interpretability: The ability to explain or present a machine learning model's reasoning in terms understandable to a human**.**

Machine Learning (ML): A subset of AI, which builds (trains) a predictive model from input data. Therefore, these AI systems are probabilistic.

Model creep, see *model drift*

Model drift: The degradation of model performance due to changes in data and relationships between input and output variables.

Neural Nets, see *Artificial Neural Networks*

Parameter: A variable of a model that the machine learning system learns on its own.

Prediction: A model's output when provided with an input example.

Privacy violation: The accessing or sharing of information without permission.

Production model: A machine learning model that has been launched into operation after being successfully trained and evaluated.

Protected variable: The features that may not be used as the basis for decisions, such as race, religion, national origin, gender, marital status, age and socioeconomic status**.**

Recommender system: A system that selects for each user a relatively small set of desirable items from a large corpus of possible options that are most likely to meet the requirements of that user.

Regression: A type of model that outputs continuous values.

Reinforcement Learning (RL): A family of algorithms that learn an optimal policy, whose goal is to maximise return when interacting with an environment.

Replication, *see explainability*

Scope creep: The case where a model expands during development to incorporate more variables and/or data, yet fails to secure consent for personal data to be used for that given purpose, even if the organisation has rightfully obtained the data in the first place.

Supervised Learning (SL): Training a model from input data and its corresponding labels.

Testing: A final, real-world check using a dataset unseen by the machine learning algorithm to confirm that it was trained effectively.

Tuning: A trial-and-error process by which some hyperparameters are changed, and the algorithm is run on the data again. Its performance is then compared with the validation set to determine which set of hyperparameters results in the most accurate model**.**

Unsupervised learning (USL): Training a model to find patterns in an unlabelled dataset.

Validation: A process used to evaluate the quality of a model using a different subset or subsets of the data, other than the training data.

X-AI, see *explainability*

References

1. Floridi, L. and J. Cowls, *A unified framework of five principles for AI in society*, in *Ethics, Governance, and Policies in Artificial Intelligence*. 2021, Springer. p. 5-17.
2. McKinsey, *McKinsey. Artificial intelligence in the United Kingdom: prospects and challenges*. 2019.
3. Taddeo, M. and L. Floridi, *How AI can be a force for good.* Science. 2018. 361(6404): p. 751--752.
4. Government Digital Service (GDS) and the Office for Artificial Intelligence (OAI), *A guide to using artificial intelligence in the public sector.* London, UK. 2019.
5. Cowls, J., et al., *A definition, benchmark and database of AI for social good initiatives.* Nature Machine Intelligence. 2021. 3(2): p. 111-115.
6. Taddeo, M. and L. Floridi, *What is data ethics?* Philosophical Transactions of the Royal Society A. 2016: p. 1-5.
7. Tsamados, A., et al., *The ethics of algorithms: key problems and solutions.* AI & Society, 2021: p. 1-16.
8. EU High Level Expert Commission, *Ethics Guidelines for Trustworthy Artificial Intelligence*. 2020.
9. OECD, *Recommendation of the Council on Artificial Intelligence*. 2019.
10. European Commission, *The Artificial Intelligence Act*. 2021.
11. Floridi, L., *Translating principles into practices of digital ethics: five risks of being unethical*, in *Ethics, Governance, and Policies in Artificial Intelligence*. 2021, Springer. p. 81-90.
12. Mökander, J. and M. Axente, *Ethics-based auditing of automated decision-making systems: intervention points and policy implications*, in *AI & Society*. 2021, Springer London.
13. Mökander, J. and L. Floridi, *Ethics-based auditing to develop trustworthy AI*, in *Minds and Machines*. 2021, Springer Netherlands. p. 2-6.
14. Mökander, J., et al., *Conformity assessments and post-market monitoring: a guide to the role of auditing in the proposed European AI regulation.* Minds and Machines, 2021: p. 1-27.
15. Holweg, M.Y., R. Younger and Y. Wen, *The reputational risks of AI.* California Management Review - Insights, 2022.
16. Whittlestone, J., et al., *The role and limits of principles in AI ethics: towards a focus on tensions*. in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 2019.
17. Mökander, J., et al., *Conformity assessments and post-market monitoring: a guide to the role of auditing in the proposed European AI regulation.* Minds and Machines, 2021: p. 1-27.
18. Hodges, C., *Ethics in business practice and regulation.* Law and Corporate Behaviour: integrating theories of regulation, enforcement, compliance and ethics, 2015: p. 1-21.
19. Ferdows, K. and A. de Meyer, *Lasting improvements in manufacturing performance: in search of a new theory.* Journal of Operations Management, 1990. 9(2): p. 168-184.
20. Holweg, M., et al., *Process theory: the principles of operations management*. 2018: Oxford University Press.
21. Goodfellow, I., Y. Bengio, and A. Courville, *Deep learning*. *Adaptive computation and machine learning*. 2016, Cambridge, Massachusetts: The MIT Press.

22. Brynjolfsson, E., D. Rock, and C. Syverson, *Artificial intelligence and the modern productivity paradox: a clash of expectations and statistics*, in *The Economics of Artificial Intelligence*, A. Agrawal, Editor. 2019. University of Chicago Press.
23. d'Alessandro, B., C. O'Neil, and T. LaGatta, *Conscientious classification: a data scientist's guide to discrimination-aware classification.* Big data, 2017. 5(2): p. 120-134.
24. IEEE, *IEEE Guide-Adoption of ISO/IEC TR 24748-1:2010 Systems and Software Engineering--Life Cycle Management-Part 1: Guide for Life Cycle Management*, in *IEEE Std 24748-1-2011*. 2011. p. 1-96.
25. Chollet, F., *Deep learning with Python*. 2017. Simon and Schuster.
26. Turilli, M., *Ethical protocols design.* Ethics and Information Technology, 2007. 9(1): p. 49-62.
27. Morley, J., et al., *From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices.* Science and Engineering Ethics, 2020. 26(4): p. 2141-2168.
28. D'Agostino, M. and M. Durante, *Introduction: the governance of algorithms.* Philosophy & Technology, 2018. 31(4): p. 499-505.
29. Raji, I.D., et al. *Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing*, in *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 2020.
30. Cath, C., et al., *Artificial intelligence and the 'good society': the US, EU, and UK approach.* Science and Engineering Ethics, 2018. 24(2): p. 505-528.
31. Floridi, L., et al., *AI4People – an ethical framework for a good AI society: opportunities, risks, principles, and recommendations.* Minds and Machines, 2018. 28(4): p. 689-707.
32. Mittelstadt, B., *Principles alone cannot guarantee ethical AI.* Nature Machine Intelligence, 2019. 1(11): p. 501-507.
33. Kahneman, D. and G. Klein, *Conditions for intuitive expertise: a failure to disagree.* American Psychologist, 2009. 64(6): p. 515-526.
34. Breck, E., et al. *The ML test score: a rubric for ML production readiness and technical debt reduction*, in *EEE International Conference on Big Data (Big Data)*. 2017. IEEE.
35. Akkiraju, R., et al. *Characterizing machine learning processes: a maturity framework*, in *Conference on Business Process Management*. 2020. Springer.
36. Van den Bergh, J. and D. Deschoolmeester, *Ethical decision making in ICT: discussing the impact of an ethical code of conduct.* Communications of the IBIMA, 2010: p. 1-10.
37. Google PAIR. *People + AI Guidebook*. 2019 May 18, 2021 [cited 10 October, 2021]; Available from: https://pair.withgoogle.com/guidebook.
38. Taddeo, M., et al., *Artificial intelligence and the climate emergency: opportunities, challenges, and recommendations.* One Earth, 2021. 4(6): p. 776-779.
39. Hapke, H. and C. Nelson, *Building machine learning pipelines*. 2020: O'Reilly Media.
40. Miller, B. and I. Record, *Justified belief in a digital age: on the epistemic implications of secret internet technologies.* Episteme, 2013. 10(2): p. 117-134.
41. Mittelstadt, B.D., et al., *The ethics of algorithms: mapping the debate.* Big Data & Society, 2016. 3(2): p. 2053951716679679.
42. Tsamados, A., et al., *The ethics of algorithms: key problems and solutions.* AI & Society, 2021.

43. Loshin, D., *Chapter 5 - Data quality and MDM*, in *Master Data Management*, D. Loshin, Editor. 2009, Morgan Kaufmann: Boston. p. 87-103.
44. Agrawal, A., J. Gans, and A. Goldfarb, *Prediction machines: the simple economics of artificial intelligence*. 2018: Harvard Business Press.
45. Hastie, T., R. Tibshirani, and J. Friedman, *Unsupervised learning*, in *The Elements of Statistical Learning*. 2009, Springer. p. 485-585.
46. Choudhury, P., E. Starr, and R. Agarwal, *Machine learning and human capital complementarities: experimental evidence on bias mitigation.* Strategic Management Journal, 2020. 41(8): p. 1381-1411.
47. Sculley, D., et al., *Machine learning: the high interest credit card of technical debt.* 2014.
48. Kaplan, R.S. and D.P. Norton, *Using the balanced scorecard as a strategic management system.* Harvard Business Review, 2007. 85(7/8): p. 150-161.
49. Kaplan, R.S. and D.R. Norton, *The balanced scorecard: measures that drive performance. (Cover story)*, in *Harvard Business Review*. 2005, Harvard Business School Publication Corp. p. 172.
50. Mitchell, M., et al. *Model cards for model reporting*. in *Proceedings of the conference on fairness, accountability, and transparency*. 2019.
51. Cunningham, W., *The WyCash portfolio management system.* ACM SIGPLAN OOPS Messenger, 1992. 4(2): p. 29-30.
52. Hutchinson, B., et al. *Towards accountability for machine learning datasets: Practices from software engineering and infrastructure*, in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 2021.
53. Partnership on AI, *Human–AI collaboration: Framework and case studies*. 2019, Partnership on AI.
54. Kaplan, R. and A. Mikes, *Managing risks: a new framework.* Harvard Business Review, 2012.
55. AI HLEG, *Ethics guidelines for trustworthy AI*. 2019, European Commission.
56. OECD, *Scoping the OECD AI principles.* 2019.
57. Chen, X., et al., *Microsoft COCO captions: data collection and evaluation server.* arXiv preprint arXiv:1504.00325, 2015.
58. Thomee, B., et al., *The new data and new challenges in multimedia research.* arXiv preprint arXiv:1503.01817, 2015. 1(8).
59. Gordon, J. and B. Van Durme, *Reporting bias and knowledge extraction.* 2013.
60. Misra, I., et al. *Seeing through the human reporting bias: visual classifiers from noisy human-centric labels*, in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
61. Mehrabi, N., et al., *A survey on bias and fairness in machine learning.* ACM Computing Surveys (CSUR), 2021. 54(6): p. 1-35.
62. Suresh, H. and J.V. Guttag, *A framework for understanding unintended consequences of machine learning.* arXiv preprint arXiv:1901.10002, 2019. 2.
63. Brewer, M.B., *In-group bias in the minimal intergroup situation: A cognitive-motivational analysis.* Psychological bulletin, 1979. 86(2): p. 307.

64. Google. *Machine Learning Glossary: Fairness*. 2021 [cited 29 November, 2021]; available from: https://developers.google.com/machine-learning/glossary/fairness.

65. Tajfel, H., et al., *An integrative theory of intergroup conflict,* in *The social psychology of intergroup relations.* 1979, Monterey, CA. p. 33-47.

66. Pedreshi, D., S. Ruggieri, and F. Turini. *Discrimination-aware data mining*, in *Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining*. 2008.

67. Clarke, K.A., *The phantom menace: Omitted variable bias in econometric research.* Conflict Management and Peace Science, 2005. 22(4): p. 341-352.

68. Mustard, D.B., *Reexamining criminal behavior: the importance of omitted variable bias.* Review of Economics and Statistics, 2003. 85(1): p. 205-211.

69. Riegg, S.K., *Causal inference and omitted variable bias in financial aid research: assessing solutions.* The Review of Higher Education, 2008. 31(3): p. 329-354.

70. Wirth, R. and J. Hipp. *CRISP-DM: Towards a standard process model for data mining*, in *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*. 2000. Springer London.

71. Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth, *The KDD process for extracting useful knowledge from volumes of data.* Communications of the ACM, 1996. 39(11): p. 27-34.

72. Amershi, S., et al. *Software engineering for machine learning: a case study*, in *2019 IEEE/ACM 41st international conference on software engineering: software engineering in practice (ICSE-SEIP)*. 2019. IEEE.

73. Silver, H. and L. Danielowski, *Fighting housing discrimination in Europe.* Housing Policy Debate, 2019. 29(5): p. 714-735.

74. Chawla, N.V., et al., *SMOTE: synthetic minority over-sampling technique.* Journal of Artificial Intelligence Research, 2002. 16: p. 321-357.

75. Fernández, A., et al., *SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary.* Journal of Artificial Intelligence Research, 2018. 61: p. 863-905.

76. Li, L., et al., *Hyperband: a novel bandit-based approach to hyperparameter optimization.* Journal of Machine Learning Research, 2017. 18(1): p. 6765-6816.

77. Snoek, J., H. Larochelle, and R.P. Adams, *Practical bayesian optimization of machine learning algorithms.* Advances in Neural Information Processing Systems, 2012. 25: p.2960-2968.

78. Bergstra, J., et al., *Algorithms for hyper-parameter optimization.* Advances in Neural Information Processing Systems, 2011. 24: p. 2546-2554.

79. Hutter, F., H.H. Hoos, and K. Leyton-Brown. *Sequential model-based optimization for general algorithm configuration*, in *International conference on learning and intelligent optimization*. 2011. Springer.

80. Liaw, R., et al., *Tune: A research platform for distributed model selection and training.* arXiv preprint arXiv:1807.05118, 2018.

81. Golovin, D., et al. *Google vizier: a service for black-box optimization*, in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 2017.

82. Thornton, C., et al. *Auto-WEKA: combined selection and hyperparameter optimization of classification algorithms*, in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2013.

83. Feurer, M., et al., *Auto-sklearn: efficient and robust automated machine learning*, in *Automated Machine Learning*. 2019, Springer, Cham. p. 113-134.

84. Vartak, M., et al. *Mistique: a system to store and query model intermediates for model diagnosis*, in *Proceedings of the 2018 International Conference on Management of Data*. 2018.

85. Hardt, M., E. Price, and N. Srebro, *Equality of opportunity in supervised learning.* Advances in Neural Information Processing Systems, 2016. 29: p. 3315-3323.

86. Guo, C., et al. *On calibration of modern neural networks*, in *International Conference on Machine Learning*. 2017. PMLR.

87. Bernstein, D., *Containers and cloud: from lxc to docker to kubernetes.* IEEE Cloud Computing, 2014. 1(3): p. 81-84.

88. Baldini, I., et al., *Serverless computing: current trends and open problems*, in *Research Advances in Cloud Computing*. 2017, Springer. p. 1-20.

89. Mittelstadt, B., et al., *The ethics of algorithms: mapping the debate.* Big Data & Society, 2016. 3(2): p. 205395171667967.

90. Tsamados, A., et al., *The ethics of algorithms: key problems and solutions.* SSRN Electronic Journal, 2020(August).

91. Mishina, Y., E.S. Block, and M.J. Mannor, *The path dependence of organizational reputation: How social judgment influences assessments of capability and character.* Strategic Management Journal, 2012. 33(5): p. 459-477.

92. ICO, *Guidance on the AI auditing framework: draft guidance for consultation.* Information Commissioner's Office, 2020.

93. PwC, *PwC Ethical AI Framework.* 2020.

94. Deloitte, *Deloitte introduces trustworthy AI framework to guide organizations in ethical application of technology*. August 26, 2020. New York.

95. Orcaa, *It's the age of the algorithm and we have arrived unprepared.* 2020.

96. IIA. *The Institute of Internal Auditors artificial intelligence auditing framework: practical applications Part A*, in *Global Perspectives and Insights*. 2017.

97. Weiss, I.R., *Auditability of software: a survey of techniques and costs.* MIS Quarterly: Management Information Systems, 1980. 4: p. 39-50.

98. Sandvig, C., et al., *An algorithmic audit.* New America Foundation, 2014(Seeta Pe \~ n a Gangadharan (ed.)): p. 6-10.

99. Sandvig, C., et al., *Auditing algorithms.* ICA 2014 Data and Discrimination Preconference, 2014: p. 1-23 , pmid = 84635968.

100. HLEG, A., *Assessment list for trustworthy AI (ALTAI)*. 2020. p. 3-33.

101. CNIL, *Privacy impact assessment - methodology*, in *Commission Nationale de l'Informatique & Libertés*. 2019. p. 400.

102. IAF, *Ethical data impact assessments and oversight models.* Information Accountability Foundation, 2019.

103. Cabrera, Á.A., et al., *FairVis: visual analytics for discovering intersectional bias in machine learning*. 2019.
104. Microsoft, *Fairlearn: a toolkit for assessing and improving fairness in AI.* 2020: p. 1-6.
105. Leslie, D., *Understanding artificial intelligence ethics and safety: a guide for the responsible design and implementation of AI systems in the public sector.* 2019: p. 97.
106. Epstein, Z., et al., *Turingbox: an experimental platform for the evaluation of AI systems.* IJCAI International Joint Conference on Artificial Intelligence, 2018. 2018-July: p. 5826-5828.
107. Bellamy, R.K.E., et al., *AI Fairness 360: an extensible toolkit for detecting and mitigating algorithmic bias.* IBM Journal of Research and Development, 2019. 63.
108. Gebru, T., et al., *Datasheets for Datasets*. CACM 2021 available at arXiv:1803.09010.
109. Floridi, L., *The ethics of information*. Oxford scholarship online. 2014, Oxford.
110. Floridi, L., *Translating principles into practices of digital ethics: five risks of being unethical.* Philosophy and Technology, 2019. 32: p. 185-193.

The authors

Professor Luciano Floridi
Professor of Sociology of Culture and Communication and Director of the European Centre for Digital Ethics, Department of Legal Studies, University of Bologna.

Professor of Philosophy and Ethics of Information, Oxford Internet Institute, University of Oxford.

Professor Matthias Holweg*
American Standard Companies Chair in Operations Management, and Director of the Oxford Artificial Intelligence Programme, Saïd Business School, University of Oxford.

Professor Mariarosaria Taddeo
Associate Professor, Oxford Internet Institute, University of Oxford.

Dstl Ethics Fellow, Alan Turing Institute.

Javier Amaya Silva
Technology and Operations Management, Saïd Business School, University of Oxford.

Jakob Mökander
Oxford Internet Institute, University of Oxford.

Dr Yuni Wen
Centre for Corporate Reputation, Saïd Business School, University of Oxford.

* Corresponding author: Professor Matthias Holweg, E: matthias.holweg@sbs.ox.ac.uk

The Centre for Digital Ethics tackles the ethical challenges posed by digital innovation. We help design a better information society: open, pluralistic, tolerant, equitable, and just. Our goal is to identify the benefits and enhance the positive opportunities of digital innovation as a force for good, and avoid or mitigate its risks and shortcomings.

Led by Luciano Floridi, the OII's Professor of Philosophy and Ethics of Information, our aim is to identify the benefits and enhance the positive opportunities of digital innovation as a force for good and avoid or mitigate its risks and shortcomings.

The Oxford Internet Institute is a multidisciplinary research and teaching department of the University of Oxford, dedicated to the social science of the Internet.

Digital connections are now embedded in almost every aspect of our daily lives, and research on individual and collective behaviour online is crucial to understanding our social, economic and political world. Our academic faculty and graduate students are drawn from many different disciplines: we believe this combined approach is essential to tackle society's 'big questions'. Together, we aim to positively shape the development of our digital world for the public good.

Saïd Business School at the University of Oxford blends the best of new and old. We are a vibrant and innovative business school, but yet deeply embedded in an 800-year-old world-class university. We create programmes and ideas that have global impact. We educate people for successful business careers, and as a community seek to tackle world-scale problems. We deliver cutting-edge programmes and ground-breaking research that transform individuals, organisations, business practice, and society. We seek to be a world-class business school community, embedded in a world-class university, tackling world-scale problems.

Saïd Business School
University of Oxford
Park End Street
Oxford, OX1 1HP
United Kingdom

www.sbs.oxford.edu