arm

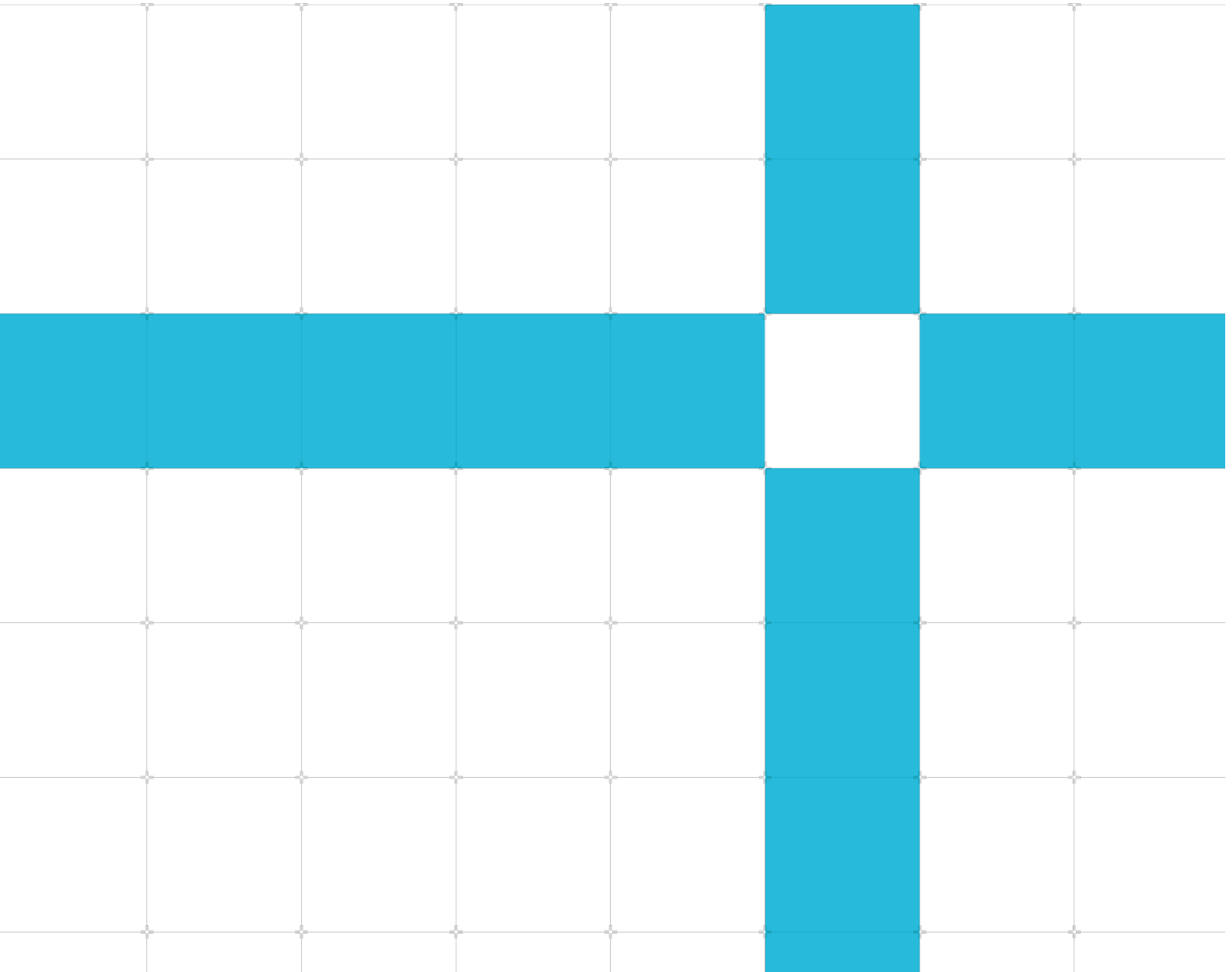
Building trustworthy AI

How a chain of assurance can build trust

Non-Confidential

Issue 0.2

Hugo Vincent, Lead Security Research Architect, Arm Research
Dominic Mulligan, Principal Research Engineer, Arm Research
Michael Lu, Director, Strategy, Security & Privacy
Stephen Pattison, VP, Public Affairs
Remy Pottier, Director, Technology Innovation
Ada Coghen-Brewster, Senior Information Developer



Building trustworthy AI

How a chain of assurance can build trust

Release information

Document history

Issue	Date	Confidentiality	Change
01	18 May 2021	Non-Confidential	-
02	29 April 2022	Non-Confidential	Updated

Arm Limited. Company 02557590 registered in England.

110 Fulbourn Road, Cambridge, England CB1 9NJ.

(LES-PRE-20349)

Non-Confidential Proprietary Notice

Arm Limited. Company 02557590 registered in England.

110 Fulbourn Road, Cambridge, England CB1 9NJ.

(LES-PRE-20349)

Confidentiality Status

This document is Non-Confidential. The right to use, copy and disclose this document may be subject to license restrictions in accordance with the terms of the agreement entered into by Arm and the party that Arm delivered this document to.

Unrestricted Access is an Arm internal classification.

Web Address

developer.arm.com

Inclusive language commitment

Arm values inclusive communities. Arm recognizes that we and our industry have used language that can be offensive. Arm strives to lead the industry and create change. We believe that this document contains no offensive language. To report offensive language in this document, email **terms@arm.com**.

Contents

1 Introduction and overview

The application of Artificial Intelligence (AI) [1] to many aspects our lives promises to be the most transformative technology trend in our lifetimes. It is crucial to the long-term development of the metaverse, which we understand as a massive computer-generated virtual world that will be deeply intertwined with our physical world. The metaverse will enable humans to live and interact in parallel or in superposition in both physical and digital worlds.

However, if AI systems are not considered trustworthy, we will miss out on all the benefits they can bring. Mass adoption of metaverse applications will have to rely on Trustworthy AI principles, just like any other digital transformation. Assessing the trustworthiness of AI systems will help avoid potential harm. Arm has been looking at trustworthy AI for the past few years and we published the Arm AI Trust Manifesto[2] describing some of the key principles that we believe should be at the heart of the debate.

Our Manifesto joined the various industry attempts in recent years to establish principles for ethical and trustworthy AI. A concerted effort by the sector to look at how to put our principles into practice will help build public trust. Some regulatory authorities are also on the point of proposing regulation. The technology sector needs to be able to show we have thought about how to put regulatory objectives into practice.

In this paper, we outline what we are calling a *chain of assurance*, which would require a company in an AI supply chain to state what ethical risks it has identified relevant to that company and state how it had addressed them.

We also look at how some emerging advances in technology can help. Our focus is on developments around security and privacy technologies, such as trusted execution environments, and how they can be used to deliver a chain of assurance, and in turn, build trustworthy AI systems.

Our ambitious exploration of these ideas is built on the fact that we have met this type of problem before, in dealing with the need to drive up security for the Internet of Things (IoT). Here we have seen how various organizations alongside regulators have shaped thinking, by offering practical proposals for putting IoT Security into practice.

These include the Platform Security Architecture (PSA) approach, of which Arm is a founder member. PSA Certified offers a detailed checklist of measures designed to help IoT device developers ensure their device is designed with security in mind right from the start. We believe that the industry should arrive at a similar point for trustworthy AI.

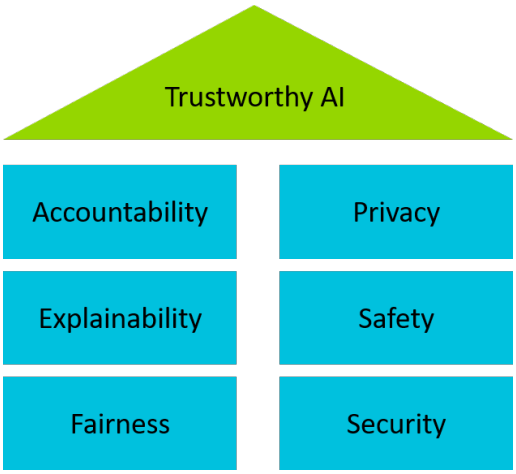
2 The landscape

It would be impossible for us to provide an overview of all the different approaches to ethical AI. The World Economic Forum suggests that over 175 organizations have proposed their own sets of ethical AI principles.[3] A quick overview of the regulatory landscape in three major legal jurisdictions is provided in Appendix A Regulators and government initiatives. Despite the variety of approaches, there is significant convergence on what the ethical guidelines for AI should be. For an AI system to be considered trustworthy, it must adhere to the following principles:

- Security
- Safety
- Privacy
- Fairness
- Explainability
- Accountability

These principles are outlined in Figure 2-1.

Figure 2-1 Building trustworthy AI



3 Chain of assurance

The core idea of a chain of assurance is that all stakeholders in the AI supply chain issue a statement describing:

- What trust-related issues relevant to their piece of IP they have considered.
- How they have addressed these issues.

This does not necessarily mean that the stakeholder has resolved all the issues listed. They may have concluded that others in the supply chain were better placed than them to do so.

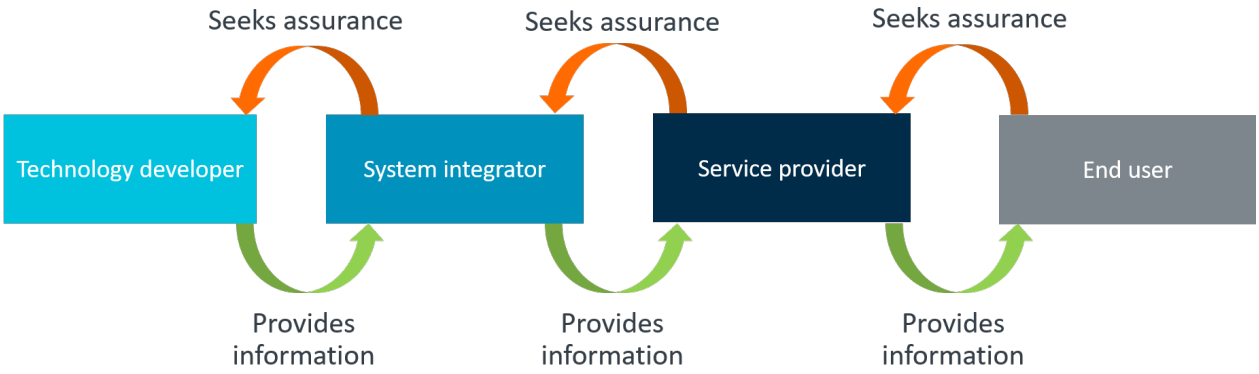
As a minimum, this would provide the company finally placing an AI service on the market with a suite of statements from the supply chain showing how potential trust issues had been tackled.

IBM floated a similar approach in 2018, noting that: 'Industries use transparent, standardized, but often not legally required documents called Supplier's Declarations of Conformity (SDoCs) to describe the lineage of a product along with the safety and performance testing it has undergone.'[4]

In the article "Towards Trustworthy AI: Mechanisms for Supporting Verifiable Claims"[5] OpenAI and others also developed the importance of verifiable claims and suggested various steps that different stakeholders in AI development can take to demonstrate responsible behavior.

We would like to build on this idea. A chain of assurance could look a bit like this.

Figure 3-1 Chain of assurance



For the sake of convenience, this drawing is a simplification: for one product, there may well be hundreds of technology developers, tens of system integrators, and a hierarchy of service providers involved. For more information about types of stakeholders we suggest for the chain of assurance, see Appendix B: Stakeholders. The chain of assurance unites all these stakeholders in their common interests. For example, for the key area of security:

- The technology developer reassures the system integrator that the technology was developed using a secure development lifecycle and in a secure environment.
- The system integrator reassures the service provider that all security requirements have been met in the integration process.

- The service provider reassures the end-user that information being stored on their service benefits from state-of-the-art security. To prove this claim, the service provider can use statements provided both by the technology developer and the system integrator.

4 Addressing the key problems

The following sections provide a deeper dive into the key areas suggested in section 2.

The chain of assurance box at the end of each section provides questions that could be asked by stakeholders, auditors, or regulators seeking assurance in the key area. These questions have been compiled drawing on the useful work of the EU's High Level Expert Group on Ethical AI[6].

4.1 Security

For security audits, verifiability of security claims continues to be the basis for assurance. It enables relying parties to make decisions about the appropriate usage of systems. Under a legislative regime or in situations where there is a high level of liability, such as potential for human injury and environmental harm, a high-level of audit and documentation is required for all systems. Using a chain of assurance approach will enable AI systems to meet those audit requirements.

We distinguish two aspects of security in relation to the implementation of trustworthy AI systems: cybersecurity in general and AI specific security.

4.1.1 Cybersecurity

The aim of cybersecurity in a more traditional sense is to implement systems that protect the confidentiality and integrity of assets, while guaranteeing a certain level of availability. Traditional cybersecurity is a constantly evolving field and there are many approaches, but at Arm we believe a security by design system is based on the four key principles of

- Analyze: Make a threat model to determine your security requirements.
- Architect: Use an established security architecture.
- Implement: Create a high-quality implementation.
- Certify: Get an independent, unbiased security evaluation of the system to create assurance.

The foundations of a secure hardware platform implementation are:

- A hardware Root of Trust that is resistant to certain physical attacks.
- Hardware-backed isolation primitives, such as the Realms of the recently announced Arm Confidential Computing Architecture, which protect against software and hardware attacks by untrusted third parties. Robust use of encryption that protects against communication attacks on data at rest and in transit.
- Secure lifecycle management to protect against supply chain attacks.

More information about secure hardware platforms can be found at the PSA (Platform Security Architecture) Certified website. For more information about using confidential computing to increase the security of AI systems, see section 5.2. We cannot stress enough the importance of platform security, especially secure lifecycle management. This lies at the heart of many of the mitigations we mentioned. It also gives us the ability to revoke or update components in the systems, whether they are models or part of the compute subsystem.

4.1.2 AI-specific security

There are also specific threats that are introduced by the development and deployment of AI systems, primarily related to the data and models inherent in these AI systems. We can categorize them in terms of attacks on:

- Data

  During the input stage to achieve a future adverse outcome, or at the inference stage, to gain knowledge of the input data used.

- Models and algorithms

  To achieve an adverse outcome, or to gain knowledge of the model/algorithm itself which was not intended.

4.1.2.1 Compromising data

- Poisoned data backdoor attacks

  This attack injects poisoned sampling data at the development stage (by model developer) or at the deployment stage (by a service provider), such that a specific unintended outcome (backdoor) can be triggered by a specific set of inputs. [7]

- Data protection and privacy risks

  For information about privacy risks, see section 4.3. One specific example of privacy risks is violation of data privacy by adversaries using member inference attacks.

- Member inference attacks

  A member inference attack attempts to establish whether a subject belongs to a specific data set through probing of APIs and running numerous inferences through machine-learning-as-a-service. [8] This presents privacy risks for private data sets, or when the data set is facial or other biometric data.

4.1.2.2 Attacks for models and algorithms

- Physical adversarial example attacks on deep learning models

  This is a well-known attack vector, where small perturbations in input can lead to a high rate of misclassification and mislead systems to potentially dangerous outcomes. [9],[10]

- Model extraction attacks

  Like member inference attacks, model extraction attacks attempt to infer properties of a model itself through manipulation of input data and output analysis.[11]

Chain of assurance questions for security
Seeking assurance: • Have the technology developer, system integrator, and device provider considered security and vulnerability? • Have the technology suppliers considered relevant trustworthy AI vulnerabilities during the design and development phase? Providing detailed information: • Does the hardware design include a hardware Root of Trust? • Does the system use secure lifecycle management to protect throughout its design, deployment, and maintenance? • What technology does the system use to ensure the integrity of data sets? • How does the system protect the security of models and algorithms?

4.2 Safety

AI systems may pose new challenges to human safety, a key concern for regulators. Human safety must be a primary consideration in the design of any AI system.[12]

We distinguish a few key areas in assessing the safety of AI technologies:

- Predictability
- Reliability
- Controllability
- Security

The general topic of reliability engineering, which ensures that products operate as intended with defined performance characteristic, without failures, under diverse though expected hostile environmental conditions, is not within scope here.

Controllability is a specific risk associated with AI systems which are designed to perform actions without human intervention, and this risk is heightened in systems with the ability to self-repair, self-improve or self-replicate.[13] In this paper, controllability is understood in terms of system failure prevention (reliability) and ensuring outcomes are as intended (predictability) and not in the general sense of ability to control and contain a generic AI system.

Security is crucial to ensure that the AI system is safe from malicious actors. For more information about security, see section 4.1.

4.2.1 Predictability

The following areas strongly affect the predictability of an AI system:

- Constraining the outputs

  An AI system is almost always implemented as part of a larger system or application, with other components or indeed other AI systems relying on its output. To achieve some measure of predictability, the outputs of an AI system must be bounded and designed to serve as input in a larger system.

- Reproducibility

  Predictability is also associated with the issue of reproducibility specific to AI system. Should the system react in the same way when inputs are equivalent? The answer to the question has implications on how the entire system responds and is a key measurement for predictability of the entire system.

- Access and availability of internal tooling and infrastructure

  Lack of access to resources used by the AI systems is cited as one of the main blocking points for predictability.[14] For predictability, the relevant resources used by the system include the data and the code base of the frameworks used, as well as the hardware version and the associated software releases.

4.2.2 Reliability

We deal specifically with the risk of unreliability in safety critical applications. To prevent loss of reliability, we need to understand potential sources of failure. The following causes of failure were highlighted in recent research.[15]

- Bad or inadequate data

  Errors introduced through bad or inadequate data at development or deployment stage can lead to differential performance to the extent that the data is not fit for purpose for certain cases.

- Shifts in environment

  Differences or shifts in environment between development and deployment can lead to again worse performance in unanticipated environments. This is where reproducibility and predictability mitigations are important considerations.

- Faulty model assumptions and/or fragile models

  Errors can be introduced both by faulty model assumptions and/or fragile models. For more information on recommendations for protecting models, see section 5.2.4.

Chain of assurance questions for safety
Seeking assurance: • Have the technology developer, system integrator, and device provider considered safety and prevention of harms? • Can this AI be approved for a high-risk or high-liability industry? Providing detailed information: • What technologies does the system use to ensure the integrity of data sets? • Does the system use attestable compute environments to ensure reproducibility? • Does the device have a manual override or fallback mechanism?

4.3 Privacy

AI systems are increasingly trained on highly sensitive personal data: both in centralized data lakes, as well as on edge devices. Many modern machine learning techniques rely on access to large data sets. The more data that the data set contains, and the more attributes that each record in the data set

possesses, the more useful a data set tends to be for machine learning purposes. These data sets pose a privacy hazard, both during the ML training phase when sensitive data is pooled together — potentially on an untrusted device — and during inference, when a trained ML model can be "probed" by a malefactor to infer information about the data set used to train the model.

What are the latest state-of-the-art technologies and best practices that can balance the tension between context-relevant personalization and society's concerns about mass surveillance and secondary use?

There are several ways to address the privacy hazards associated with the large data sets used by machine learning.

4.3.1 Anonymization

One approach is to modify the data set or limit the types of questions that can be asked about the underlying data in the data set, using anonymization techniques.

Naïve data anonymization

Naïvely, data sets can be anonymized by removing attributes that appear to be particularly sensitive – such as names, addresses, dates of birth, and so on – from their records. However, several kinds of privacy attack can be used to reconstruct or deanonymize data from data sets that were manipulated using careless anonymization techniques:

- Linkage attacks can be used to link records in an anonymized data set with records appearing in another public data set. These attacks can be surprisingly powerful: in one infamous example, a linkage attack was used to reveal the Governor of Massachusetts' health records after an anonymized medical data set was linked against freely available voter registration data[16]. In addition, background knowledge can also be used to deanonymize data. For example, knowing that heart attacks occur at a reduced rate in Japanese patients, compared to other nationalities, can be used to narrow the range of values of an attribute in medical data sets[17].
- Differencing attacks use carefully constructed sets of complementary queries over data sets – even very large ones – to infer the attributes of private records. By issuing the complementary pair of queries "how many people in this database are known to have cancer?" and "how many people in this database, not named John Smith, have cancer?", an attacker can infer information about the health of John Smith without having to *directly* query for that information.

Query auditing techniques

Query auditing techniques use explicit checks on data queries to try to gauge if the results of those queries can cause a privacy breach, before being applied to the data set. If a query can cause a privacy breach, it is blocked. Query auditing techniques would appear to be a good first defence against certain types of differencing attack. Unfortunately, this is not the case, as refusing to process a query in the context of a series of previous queries may itself reveal sensitive information about the underlying data set. Moreover, depending on the expressiveness of the underlying query language, detecting a potential breach of privacy from a series of queries may not even be computationally feasible.

Given the problems with naive data anonymization, computer scientists have sought to give a precise definition of data privacy and an associated framework within which informative queries over data sets can be made without necessarily sacrificing privacy.

k-anonymity

Early attempts at providing a framework within which data privacy could be evaluated such as *k-anonymity*[18] – and its many refinements such as *t-closeness*[19] and *l-diversity*[20] – focussed on formalizing the naive idea of anonymizing a data set by removing record attributes, as discussed above. Intuitively speaking, a data set has the *k*-anonymity property – where *k* is a *privacy parameter* – if the record of any one individual appearing in the data set cannot be distinguished from the records of any of *k – 1* other individuals also appearing in the data set. This property therefore entails that any one individual in the data set has a form of "plausible deniability" with respect to the results of queries over the data set. Data sets can be manipulated by removing attributes, or making the range of values that an attribute can take on to be less precise, so that the data set eventually satisfies the *k*-anonymity property for some desired, *k*. Unfortunately, *k*-anonymity and variants are still subject to a range of privacy attacks.

Differential privacy

Differential privacy's[21] modern form was perfected by cryptographers and makes use of a modified form of a central concept in the theory of cryptography: indistinguishability.[22] Whereas techniques such as *k*-anonymity, try to protect sensitive data by manipulating a data set, differential privacy focusses on the queries, or generalized "algorithms", that can be made about or computed over a data set.

The central observation behind differential privacy is inherently intuitive: an individual's privacy cannot be compromised by an inadvertent release of data from a data set if that data set does not contain any data related to that individual. As a result, a query over a data set is differentially private if it is indistinguishable to an external observer whether the query was computed over a data set containing an individual's data, or over a data set where that data had been removed. Indistinguishability is achieved by adding carefully chosen random noise to the output of a query. Naturally, the amount of noise to add to the output of a query is a function of the data set itself. For example, if a data set contains information about only a single person, then the amount of statistical noise needed to achieve this indistinguishability property is necessarily much greater than is needed to mask the inclusion or exclusion of an individual's data in a query over a data set containing data about all 500 million Europeans.

Differential privacy is now routinely deployed as a means of guaranteeing the privacy of individuals appearing in large data sets. For example, both Apple[23] and Microsoft[24] use variants of differential privacy to anonymise telemetry information originating from devices running their operating systems, and the US Census Bureau also uses differential privacy when aggregating population-wide statistics[25].

However, despite real-world deployment, differential privacy is *not* a panacea and does not guarantee perfect privacy, but merely places an upper-bound on the amount of information that leaks from a query over a data set. Moreover, significant amounts of noise may be required to obtain the indistinguishability property, making differential privacy inappropriate for some uses.

4.3.2 Federated learning

An alternative approach to addressing the privacy hazards associated with the large data sets associated with machine learning is to avoid collecting large pools of potentially sensitive data in one centralized data set in the first place. Federated learning[26] is a distributed machine learning technique wherein a collection of nodes – for example mobile phones, tablets, or edge devices – each possessing their own private data set, co-operate to build a combined machine learning model without explicitly exchanging records from their respective private data sets. Instead, each node trains a local model on their respective private data sets. These models are then combined, either by a central server, or in a decentralized fashion, to produce an aggregate machine learning model. Note that this aggregate model is obtained without any private data set, stored on each node, ever leaving that node.

4.3.3 Cryptographic techniques

Several cryptographic techniques can be used to guard privacy. Homomorphic encryption schemes allow computations to take place directly on encrypted ciphertexts, without requiring that the data is first decrypted. Using this technique, data can be freely shared with untrusted third parties for processing without sharing the data itself, or the results of the data processing thereafter. In the machine learning context, private data originating from a device can be encrypted and transmitted to a central server, where inference takes place using pre-trained models, and the result thereafter transmitted back to the originating device without any private data being revealed. Secure Multiparty Computations[27] allow a group of distrusting individuals to jointly compute a function over their private data sets, without revealing those data sets to each other. Many protocols for collaborative, privacy-preserving machine learning have been developed by cryptographers[28].

4.3.4 Hardware-based trusted execution environments

Hardware-based Trusted Execution Environments (or TEEs) can be used to build systems that address many of the same use-cases as the cryptographic techniques that we surveyed above, namely the protection of data whilst in-use, and the protection of data when it is pooled amongst a group of mistrusting individuals as a means of computing a joint function over that combined data. Naturally, the use of TEEs for this purpose has both disadvantages and advantages. To deploy software within a TEE, the user needs to first establish that the TEE is trustworthy. This typically involves an *attestation protocol*. The attestation protocol is used to verify the provenance of the hardware platform and supporting firmware, which need to be trusted for the TEE to meet its security guarantees. The hardware, firmware, and attestation protocol need to be explicitly trusted to enable use of TEEs. In contrast, with cryptographic techniques, one must only trust the correctness of the design and implementation of the underlying cryptographic primitive. To their advantage, systems built around TEEs can be more flexible, being easier to deploy and configure, and easier to understand, design, and program for programmers who are not domain experts in applied cryptography. Moreover, the use of hardware-based approaches can offer significant performance benefits — being often orders of magnitudes faster and capable of handling much larger data sets, running at near native speeds – than comparable cryptographic techniques.

Chain of assurance questions for privacy
Seeking assurance: • How does this system address consumer concerns associated with data governance and security? Providing detailed information: • If possible, does the system use edge processing to protect data privacy? • What techniques are being used to provide data privacy?

4.4 Fairness and bias

The issue of what constitutes unfair bias will inevitably be context specific and will involve many wider factors – such as local culture or societal attitudes.

This white paper explores the broad 'classes of bias'. It looks at how and where bias is likely to enter and how classes of bias as the first step towards identifying the sort of key questions AI developers should consider.

The key different classes of bias identified in this paper are:

- Development and training bias
- Sample/data bias
- Outcome bias

4.4.1 Development and training bias

The strategic or business intent should be properly – and fairly —reflected in the AI model development. The developer has a key role of reflecting the goal and objectives of the business and its strategic intent into a set of attributes that will then be used for AI training and inference. The choice of which attributes should be included is a major source of possible bias.

4.4.2 Sample/data bias

Bias can enter the data set as a result of the distribution of the sample data or of the character of the samples themselves.

Bias due to the data itself has been the focus of attention in the debate so far. There is also a risk that bias may relate to data access. For example, certain private data might be excluded from certain processes, which, had the data been accessible to the system, could have helped remove certain bias. For recommendations on how to protect the integrity of a data set, see section 5.1. By extension, being able to get access to certain class of data (for example, private data) could be the only way to verify if an AI is unbiased.

4.4.3 Outcome bias

Two individuals with similar characteristics with respect of metrics defined for a particular task should get a similar outcome. But even if an outcome bias is, at its root, linked to an implementation, training or data sample bias, reaching the state of outcome bias-free is inherently not possible due to

technical inaccuracy of AI systems. If we want to ensure all groups and individuals are treated the same way, then what the algorithm developers must ensure is that the probabilities of a false positive and a false negative for different groups are equal.

4.4.4 Other biases

For example, bias can be introduced in a system using federated learning due to the nature of the learning process. In federated learning there are multiple possible causes of bias: bias may be introduced by the sampling of parties and how they are query (e.g. network availability may introduce bias on how each party contributes), a model is trained on a smaller, specific set of data thus the data source may be strongly heterogenous (for example due to geo-location of the different parties) ; the fusion algorithm depending on how it weights the contribution from different parties may further amplify or introduce bias. The complexity is in the integration of model that has been trained using heterogenous data.

Chain of assurance questions for fairness and bias
Seeking assurance: • Does the system treat all groups and individuals the same way? Providing detailed information: • Has the system integrator considered possible bias in the collection of data? • Has the technology developer considered possible bias in the implementation of the data? • Has the user of the system considered possible bias in the interpretation and use of the AI based recommendation system?

4.5 Explainability

AI systems exhibit different levels of explainability: some can effectively introspect and explain why the decisions that they make were made, others less so. If an AI system's accuracy and decision-making process cannot be understood by a human being, then it is hard to assess potential risk for high-liability industries. What evidence or documentation, in the various stages of problem definition, design and development stages, enables better interpretation?

There is a popular misconception that machine learning models are necessarily inscrutable black boxes. However, several classes of machine learning model can in fact be examined introspectively. They can explain their reasoning in a human-digestible form. Decision trees and other rule-based algorithms are the classic examples, producing machine learning models that are cascading chains of "if-then" rules.

The field of eXplainable AI (XAI) is steadily developing. This introduces new machine learning algorithms for which model introspection and explainability is a first-class concern.[29] [30]

In a machine learning context, explainability is important for several reasons: it can be used to fully document the software engineering process, deduce the data and the training regimen used during model learning, and helps in evaluating the overall system performance. Moreover, a machine learning model that can explain its own reasoning is much easier to audit for compliance with relevant regulations, and, if the output of a machine learning model leads to bad or unwarranted outcomes in the real-world, an explainable model can be used to pinpoint the reasoning that led to these outcomes, which can then be modified and fixed.

However, despite recent advances in XAI, it is the case that many state-of-the-art machine learning techniques do in fact act as impenetrable black boxes to external observers. This may be because the training algorithm was not designed with model explainability in mind. Alternatively, an algorithm could in principle be designed with explainability in mind, but the size of any learned model is so gargantuan, or otherwise complex, that the chances of comprehension of any explanation by a human are slim. Many modern Deep Learning techniques, such as Convolutional Neural Networks (CNNs)[31], Long short-term memory Networks (LSTM)[32], and similar, fall into this pattern, and unfortunately these techniques also represent the state-of-the-art in several application areas of modern machine learning. We must therefore recognize that, for the near future, there will be different levels of explainability for different machine learning systems. Careful analysis is needed to understand what the appropriate level of explainability is on an application-by-application basis, and the use of explainable models should be preferred where this is appropriate and possible.

Chain of assurance questions for explainability
Seeking assurance: • How did the AI system give this outcome, and what is the reasoning behind the decision? • Why can this AI system be approved for a high-risk or high-liability industry? Providing detailed information: • Has the system integrator or service provider considered the explainability requirements for the use case? • Has the system integrator or service provider considered the model chosen in the context of explainability requirements?

4.6 Accountability

The development and deployment of AI systems inevitably involves many different subsystems and actors working together to achieve a common goal. Often these subsystems may be designed and engineered by different teams within the same company, or even different companies. How can traceable logs of steps and decisions taken in the different stages of problem definition, design, development, deployment, and operation be designed and logged to help establish liability and remediation?

During the design, all stakeholders at each level should be able to explain and eventually justify their design and development decision to other system design stakeholders. They should also be able to explore scenarios on how their decisions on design and development could ultimately impact the systems. This will help improve the global systems and help define the responsibilities of those involve.

After deployment, each element at all levels should be able to tell what happened, and why a decision (good or bad) was made. Tamper-resistant immutable logs, sometimes called append-only logs, are one such approach to solving this problem. Systems can be designed so that details of all decisions made by the system, including the details of the appropriate inputs which led to the decision being made, are appended to a tamper-resistant immutable log. If an AI system goes awry, these logs can be used by investigators as something akin to aircraft flight data recorders, or “black boxes”, to trace through the series of decisions made by the system, and the stimuli that led to those decisions being made.

Chain of assurance questions for accountability
Seeking assurance: • Who is accountable for which part of an AI based systems or product, and ultimately the AI decision or outcome? • How is this investigated if something goes wrong? Providing detailed information: • Has there been maximum transparency at all levels of the design, which includes record of how the intelligent system operates?

5 Recommendations

For each trustworthy AI principle, we outlined the challenges facing its proper implementation and suggest some questions that stakeholders should be able to answer. In this section, we suggest some possible recommendations to counter these challenges and questions. We also welcome all other suggestions and recommendations from the industry that would form a better and more useful chain of assurance for Trustworthy AI.

Most of the recommendations discussed in this paper fall within the following groups:

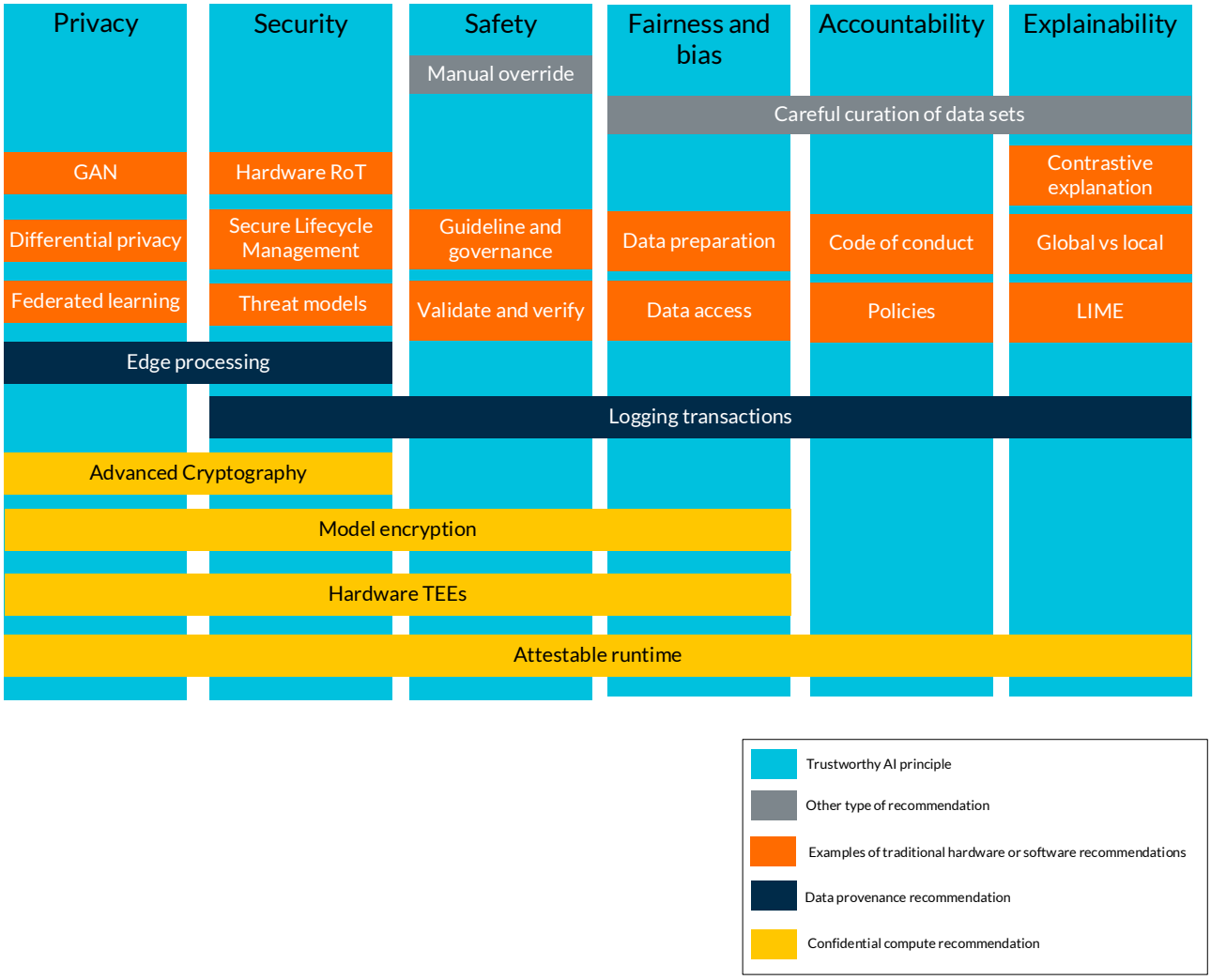
- Cybersecurity
- Data provenance
- Confidential computing

Traditional cybersecurity was discussed earlier in section 4.1.1. More information about hardware security can also be found at the PSA (Platform Security Architecture) Certified website. In this section we will discuss data privacy confidential computing and recommendations in more detail.

The recommendations we give differ: some are well-recognized solutions (such as hardware TEEs), some require more research (for example, developing unbiased data sets). Recommendations for different principles often overlap — especially for security, safety, and privacy.

Figure 5-1 shows which recommendations apply to which principles.

Figure 5-1 Trustworthy AI principles and recommendations



5.1 Data Provenance

Provenance (also sometimes referred to as lineage) metadata describes the modification history of a data set or of the origin and transformation history of data ultimately derived from a data set.

Traditionally, pools of data were centralized and assumed to be under the management of trustworthy authorities–dedicated database administrators, for instance, who restricted who could add data to the data set, and in what form. Under this assumption, it was also reasonable to assume that the data contained within any such data set was consistent, reliable, and had been vetted before being added to the data set, with trust in the content of the data set being largely implicit and unstated[33]. Today, these old assumptions no longer apply, for a variety of reasons, but notably because the rise of the Internet has led to an explosion in data creation and synthesis of data sets: data is now constantly created and modified, giving rise to potentially huge, decentralized data sets with few integrity or well-formedness guarantees.

The move away from trustworthy, centralized data sources to untrustworthy, decentralized data sources, and the increase in the size of modern data sets, spurred the development of provenance tracking techniques. A significant body of work in this area has been carried out by the database research community, who aimed to answer questions about transformations of data by relating the inputs and outputs of these transformations[34]. In particular, the database research community were motivated to study data to understand the *why*, *how*, and *where* of data:

- Given the output of some data transformation or query, can we identify inputs to the transformation that explain *why* the output was produced?
- Can we track how input data was transformed to demonstrate *how* a particular output was obtained?
- Can we identify which data sets, or which records from a particular data set, contributed to the output of some query or data transformation? That is, can we identify *where* the inputs that contributed to an output originated?

Being able to answer these questions about data sets is also useful in a machine learning and AI context, too, especially during the training of ML models from data sets. For example, by understanding why a machine learning model produced an answer, which data sets the machine learning model was trained upon, and how the answer was produced, we are better able to debug the model when it produces unfavourable results, and better able to pinpoint sources of bias when they are revealed.

Further, there is a well-attested reproducibility crisis in the machine learning community, wherein a large percentage of research papers presenting new machine learning algorithms and techniques cannot be reproduced by third parties. Provenance tracking has therefore been seen as a potential defence against this reproducibility crisis, wherein the integration of provenance tracking, and provenance metadata, into machine learning pipelines makes it easier to pinpoint exactly what data set was used in an experiment, how it was transformed, and so on, thereby making it easier to replicate. If machine learning models start to be regularly deployed in regulated domains, such as the automotive, aerospace, or medical domains, then this reproducibility then becomes vital.

5.1.1 Logging transactions

To guarantee the quality and integrity of data according to data provenance recommendations, one must log both the integrity of data sets as well precisely what has been done with each piece of data, and why. This log will allow for auditing as well as potential accident diagnosis in the development, integration, and deployment stages.

One must also log transactions to mitigate against membership inference attacks in the deployment stage. Transactions must be logged to detect and deter bad actors in the system. [35]

5.2 Confidential Computing

Confidential Computing refers to a series of techniques by which a computation, or inputs to that computation are protected from untrusted onlookers, with these onlookers neither able to observe nor interfere with the computation. Cryptographic or hardware-based Trusted Execution Environments (TEEs) may be used to implement these protected computations, or a mixture of the two. In this section, we focus on the use of TEEs in Confidential Computing. Notably, many of the

facets of Trustworthy AI, previously discussed, can be implemented using trusted hardware. For example, TEEs can be used as a mechanism for ensuring data privacy, owing to the strong confidentiality and integrity guarantees that they provide to loaded data and software. Moreover, TEEs and associated remote attestation procedures, explained further below, can be used as primitives when implementing a provenance chain for data sets.

Several TEE technologies are now available for commodity hardware, including Arm TrustZone[36], Intel's Software Guard Extensions (SGX)[37] AMD's Secure Encrypted Virtualization[38], AWS Nitro Enclaves.[39] Arm has also recently announced the Realm Management Extensions, part of Arm's Confidential Compute Architecture (CCA). Whilst each of these different isolation technologies have peculiarities of their own, some commonalities can be identified:

- Strong isolation against a privileged attacker. All the technologies surveyed above attempt to provide strong integrity and confidentiality guarantees for data and code against a privileged attacker. Arm TrustZone conceptually provides two "worlds" – Secure and Non-secure – with memory addresses tagged with their originating world. This insulates code and data residing in the Secure world from privileged code, even the operating system or hypervisor, executing in the Non-secure world. Similarly, Intel SGX provides a *Secure Enclave* which protects code and data from privileged code, including the operating system, executing on the same machine. Additionally, AMD SEV protected virtual machines, Arm CCA Realms, and Intel SGX Secure Enclaves are backed by integrity-protected encrypted memory, providing some defence against even physical attackers.
- Support for Remote Attestation. Remote attestation protocols are cryptographic methods through which a device can authenticate its hardware and software configuration to a third-party. Intuitively, remote attestation protocols allow a sceptical challenger to obtain compelling cryptographic proof, via an *attestation token*, that a device is configured in a particular way — for example, that a certain piece of software known to the challenger is installed on the device, or the device has known good configuration options set.
- Small Trusted Computing Base. The technologies above aim to reduce the amount of code that is included in the Trusted Computing Base (TCB). Notably, this includes moving the "rich" operating system and other privileged system code, such as a hypervisor, out of the Trusted Computing Base.

Numerous start-ups, and small businesses, are already offering privacy-preserving compute platforms built around TEEs, such as Cosmian[40] in France, Decentriq[41] in Switzerland, IoTeX in the United States[42], Scalys[43], in the Netherlands, SCONE[44] in Germany, amongst many others. Moreover, established companies, including major cloud providers like Microsoft Azure[45] and Amazon's AWS[46], are now also offering access to TEEs to their customers, and major financial institutions such as Ant Financial[47] in China and JP Morgan Chase in the United States are exploring the use of Strong Isolation technology to protect customer's data.[48] Confidential computing is key to implementing the key areas of security, safety, and privacy in the Trustworthy AI chain of assurance.

5.2.1 Hardware-based trusted execution environments

Using TEEs with attestable properties enables transparency and the reproduction of the compute environment used during both development as well as deployment stages. TEEs can also be used to protect the confidentiality and integrity of sensitive data sets and machine learning models.

In situations where there is a necessity to pool data or make use of potentially untrusted third-party devices to host a computation, and where advanced cryptographic techniques are currently inapplicable, we also recommend adopting the use of hardware TEEs, to protect computations on

user data. Distributed systems built around TEEs are a pragmatic solution to providing users with strong privacy guarantees. Whilst their security and privacy guarantees fall short of pure cryptography, they are now widely deployed, efficient, and easier to use for the average programmer than cryptographic approaches.

5.2.2 Edge processing

Use edge processing where possible to minimize systemic risk of data pooling. In some distributed systems designs, where data is to be pooled in a central location or shared with untrusted third parties, it may be possible to carefully design the system so that raw data never leaves a user's device thereby providing strong data privacy guarantees. This is the case with federated learning discussed in section 4.3.2. We recommend that designers of data-intensive distributed systems consider ways to limit the amount of data pooled in centralized services, and consider moving more compute onto a user's device, where this is possible.

5.2.3 Attestable runtime

The ability to report, monitor, and correlate potential errors during operation is also a critical component of reliability. Point-wise reliability, also known as real-time anomaly detection, is an increasingly important approach in the prevention as well as logging of catastrophic failures. Usage of computing infrastructure with attestable properties will reduce potential attacks on models during use. Remote attestation and runtime measurements would enable transparency and more granular monitoring of the operational environment used during both development as well as deployment stages.

5.2.4 Model encryption and decryption.

As more models are deployed to the edge, confidentiality of the model is achieved not only through executing the model in an isolated environment. Model encryption and decryption on a per device basis in the deployment stage are becoming increasingly important,[49] especially to address the model extraction attacks that was mentioned earlier in section 4.1.

5.2.5 Advanced cryptographic techniques.

Given their strong security and privacy guarantees, use advanced cryptographic techniques for centralized data processing where viable. In situations where there is a necessity to pool data in a centralized location, or where computations require more computational power than is available on a user's device, it may make sense to consider deploying advanced cryptographic techniques such as Homomorphic Encryption and Secure Multi-party Computations. Whilst these have long been deemed too inefficient for widespread industrial adoption, many techniques are now reaching a state of maturity wherein they can be deployed profitably in restricted situations.

5.3 Manual override or fallback mechanism

The ability to fail gracefully, to have a fallback which can be relied on in event of failure or unanticipated situations is critically important as a last resort if failure occurs.

Where possible, human intervention should be asked for, otherwise there needs to be a deterministic form of a rule-based (instead of non-deterministic algorithm) application or component that can take

over certain functions, which is in a verified good state. A fallback mechanism needs to be deterministic in nature, and an attestable signed version of the compute environment including certified firmware would normally suffice to ensure reproducibility and the foundation to implement as a fallback mechanism.

5.4 Careful selection of data sets

Engineers and data scientists should keep in mind that machine learning models trained on biased data sets may produce inequitable outcomes once deployed. As a result, they must ensure that sources of potential bias are eliminated from their training data sets, for example by ensuring that the demographics of training data sets used to produce computer vision models are reflective of wider society. Coupling carefully considered and curated training data sets, with a chain of provenance, allows designers to further pinpoint the source of any observed inequity.

6 Call to action

In conclusion, the advantages of developing a chain of assurance would help the tech sector should build trust in AI in order to build the gateway to the metaverse.[50] Whatever the precise shape of regulatory approaches, it is quite likely that such a scheme could help play a key role in assisting companies fully meet any future regulatory requirements. But even without the stimulus of emerging regulatory interest, a chain of assurances would be an important step towards building trust. We should not, as a sector, sit back and wait for regulation, before acting ourselves.

In this paper we have set out some of the key considerations: the major issues to be addressed, aims which need to be covered, and how and by whom.

But we recognise that key aspects of a chain of assurance scheme remain unresolved.

First, for it to be successful there needs to be some standardisation in the way such assurances are given to ensure that the assurances cover the same ground and address the right issues.

Second, we need to decide whether the assurance will be in the form a self-declaration by a company of what it has done, or whether there will be third-party verification or 'endorsement'. Both options may run simultaneously and be respectively appropriate for different use cases with the higher risk use cases aiming for third party involvement.

Most importantly we need a critical mass of companies interested in exploring these ideas to work together. Arm's success in launching Platform Security Architecture (PSA) which describes how companies can confirm they have addressed IoT security issues, shows that this kind of approach can work. As first step, Arm will use this paper to engage our partners all over the world and others, such as trade associations, in taking forward this concept for Trustworthy AI.

Appendix A Regulators and government initiatives

In this section we take a brief look at the principles and the frameworks on which various jurisdictions are trying to formulate their approach to promoting trustworthy AI. We provide a quick overview of the regulatory approaches of the EU, the US, and China

A.1 European Union

The European Commission published its wide-ranging proposals for AI regulation in April 2021. In summary, these proposals would prohibit certain uses of AI, like AI based 'social scoring', real time remote use of biometric recognition for law enforcement in public spaces (with some exceptions), and AI which is aimed at 'manipulating behavior' to limit free will. Other AI use cases are divided into

- High risk AI
  - High risk AI would require prior assessment of conformity in areas like privacy protection, human in the loop etc.
  - Examples: High risk includes transport, education, recruitment, credit scoring (and the list could be extended).
- Limited risk AI
  - The customer needs simply to be informed about the use of AI,
  - Examples: Chatbots
- Minimal risk AI
  - There are no formal requirements.
  - Examples: Video games

Earlier, the Commission had included the EU High-Level Expert Group on Artificial Intelligence (AI-HLEG). This concluded that trustworthy AI should be[51], :

- Lawful, complying with all applicable laws and regulations.
- Ethical, ensuring adherence to ethical principles and values.
- Robust, both from a technical and social perspective.

The core relevant ethical principles for the AI-HLEG were: respect for human autonomy, prevention of harm, fairness and explicability. To translate these concepts into practice, the EU's approach is likely to focus on :

- Human agency and oversight
- Technical robustness and safety
- Privacy and data governance

- Transparency
- Diversity, non-discrimination, and fairness
- Societal and environmental wellbeing
- Accountability

A.2 United States:

Federal Government thinking on what role government action could play is less focused on regulation than in the EU. But the Government has not excluded regulatory interventions where needed and acknowledges the importance of building trustworthy AI. In 2019 the President signed an Executive Order on AI which stressed the importance of standards including for building trust. Further, in 2020 the US Congress passed the National AI Initiative Act[52], which among other things created the "National AI Initiative Office" to coordinate US government R&D and policy.

The National Institute for Standards and Technology, or "NIST", which performs work that often shapes policy in these technical areas, has been looking at AI Trustworthiness. NIST is looking into how trust can be increased through development and adoption of strong technical and non-technical standards in the areas of:

- Accuracy
- Explainability
- Resiliency
- Safety
- Reliability
- Objectivity
- Security

It is expected that recommendations for policy in those areas will be made in due course. NIST notes that significant standards work is being done in many of these areas, but that these standards will need to be revised and updated as the technology advances.

One of the focus areas NIST has discussed in various workshops that it has convened on AI, is to work on development of technical standards and related tools in support of reliable, robust, and trustworthy systems that use AI technologies, specifically:

- Data sets in standardized formats, including metadata for training, validation, and testing of AI systems
- Tools for capturing and representing knowledge and reasoning in AI systems
- Fully documented use cases
- Benchmarks
- Testing methodologies
- Metrics to quantifiably measure and characterize AI technologies, including but not limited to aspects of hardware (at device/circuit/system level), trustworthiness (e.g., accuracy, explainability, safety, reliability, objectivity, and security) etc.
- AI testbeds

NIST are also in the process of developing an AI risk management framework aimed at better managing potential risks to individuals, organizations and society that could result from broader use of AI. [53] Further, the US federal government have created a central repository at **ai.gov** for AI related activity.

A.3 China

The Chinese government is committed to the development of China's AI industry and has been working on AI standards and ethical guidelines. Trustworthy AI, responsible AI, AI ethics are a key part of their discussions.

In 2017, China's State Council published the *New Generation AI Development Plan*. [54] This set a strategic goal of drafting an initial approach to laws, regulations, and ethical norms, related to AI. The aim is to have more comprehensive frameworks in place by 2030. Following this, in 2019, the Ministry of Science and Technology (MoST) issued *Development of Responsible AI: A New Generation of AI Governance Principles*. [55] The AI Governance Principles provide a framework and action guidelines for AI governance, aiming to "ensure that AI is safe/secure, reliable, and controllable" The following 8 principles of AI governance are proposed:

- Harmony and friendliness
- Fairness and justice
- Inclusiveness and sharing
- Respect for privacy
- Security and controllability
- Shared responsibility
- Open cooperation
- Agile governance

The New Generation AI Governance Expert Committee was established by the MoST in 2019 to research policy recommendations for AI governance and identify areas for international cooperation[56]. Issues of focus include data monopoly, algorithm bias, abusive use of intelligence, deep fake, data poisoning, privacy protection, ethical norms, and inequality. The Committee released a "Code of Ethics" for AI development titled the *New Generation Artificial Intelligence Ethics Specifications* in September 2021. The document outlines six fundamental ethical principles for implementing and using AI technologies in society: (1) improving human welfare, (2) promoting fairness and justice, (3) protecting privacy and security, (4) ensuring controllability and trustworthiness, (5) enhancing responsibility and (6) improving ethical literacy.

  - Under the principle of protecting privacy and security, the document states that AI users must be informed of how their data is being handled and must consent to or reject the usage of AI systems. Personal data needs to be held in accordance with "the principles of lawfulness, fairness, necessity, and integrity." The document also requires the security and transparency of the R&D and application of AI technologies.
  - Under the principle of "ensuring controllability and trustworthiness, " the document stresses that humans must "have full autonomous decision-making power and the right to choose whether to accept the services provided by AI and the right to withdraw from the interaction with an AI system at any time."

To complement the policy work, China's standards organizations work on proposals to guide and standardize the behavior of AI practitioners. TC260 (National Information Security Standardization Technical Committee), the main body for drafting the technical standards for information security, produced a report in Jan 2021, *Cybersecurity Standard Practice Guide-Guidelines for Ethical Security Risk Prevention of Artificial Intelligence*, which provides guidelines for artificial intelligence ethics and ethical issues throughout the technology lifecycle, including research and development, design and manufacturing, deployment and application and other related activities.[57] Based on the Guidelines, in August 2021, TC260 drafted the national recommended standard, *Information security technology-Assessment specification for Machine learning algorithms*, which is available for public comments. The draft standard describes the security requirements for machine learning algorithms throughout their lifecycle. Confidentiality and privacy are included in the security assessment criteria.

The MIIT-affiliated China Electronics Standardization Association released its group standard *Information Technology - Artificial Intelligence - Risk Assessment Model* for public consultation in July 2021. The risk assessment model proposed by the group standard has three risk factors – technical risk, application risk, and management risk. The technical risk is further broken down into data risk, algorithm model risk, and systematic risk.

In the same month, at the World Artificial Intelligence Conference (WAIC) in Shanghai, the Shanghai municipal government announced the establishment of the Shanghai Municipal AI Standardization Technical Committee. Standards related to information security and ethics are among the priorities of the committee's near-term plan. In addition, another MIIT-affiliated institute China Electronics Standardization Institute (CESI) stated in the *AI Standardization White Paper (2021)* that it plans to formulate a standard for the technical requirements of machine learning systems for privacy protection.

In addition to government activities, there are also several Chinese state-affiliated industry associations and think tanks focusing on guidelines for trustworthy AI development:

- Beijing Academy of Artificial Intelligence (BAAI) [58]: In 2019, BAAI released the "Beijing AI Principles" for the "research, development, use, governance, and long-term planning of AI." [59] "Be Ethical" is listed as one of the total 7 principles for the research and development of AI, calling for ethical design approaches to make the AI systems trustworthy, i.e., "making the system as fair as possible, reducing discrimination and biases, improving its transparency, explainability, and predictability, and making the system more traceable, auditable and accountable, etc." Last year, BAAI released a report on AI governance and ethics, advocating the "Beijing AI Principles" in R&D, use and governance.[7]
- China's Artificial Intelligence Industry Alliance (AIIA)[lvi]: In 2019, AIIA released a draft "joint pledge" on, among other things, the principles of secure and trustworthy AI, transparency and explainability, privacy protection, clear responsibilities, and diversity and inclusiveness. [60] A year later, AIIA published *Trustworthy AI Operation Guidelines (V0.5)*, providing a practical guidance based on the principles to execute the trustworthy AI requirements via an "ethics by design" approach. Soon after that, AIIA published *Management Measures for Trusted AI Demonstration Zone*, planning to start pilot programs in China to promote Trustworthy AI.[61] In July 2021, AIIA announced the establishment of the AI Governance and Trustworthiness Committee. To date, AIIA's efforts related to trustworthy AI mainly include the trustworthiness assessment of AI applications such as face recognition systems and RPA.
- Tsinghua University: Tsinghua University founded the Institute for AI International Governance (AIIG) in April 2020. The former Vice Minister of Ministry of Foreign Affairs, Ms. Ying Fu, is the honorary Institute President. Over the past year, AIIG established its first academic committee that consists of 11 top-level scholars at home and abroad. It also organized dozens of workshops,

high-level conferences, and enterprise visits. Currently, it is undertaking research projects on AI governance commissioned by both the Chinese government and enterprises.

Appendix B Stakeholders

B.1 Technology Developer

An organisation or individual involved in the development of technology suitable for multiple applications.

Examples of technology developers:

- Hardware manufacturers designing FPGA or specialised chips used in inferencing or training
- Data aggregators obtaining or producing curated and metadata-tagged data sets
- Neural Network model designers
- Algorithm developers producing new mathematical methods for optimising training/inferencing, or to improve accuracy of models

B.2 System integrator

An organisation or individual involved in designing and producing systems or products that is tailored for market and sector specific applications.

Examples of system integrators:

- Profiling and recommendation engine to suggest new movies and content or advertising
- Facial Recognition CCTV system manufacturers
- Credit scoring for loan applicants
- Imaging based Cancer Diagnostic system
- Robotics systems

B.3 Service Provider

An organisation or individual providing end-user facing services or products.

Examples of service providers:

- Video streaming service providing “watch next” and/or advertising to increase engagement and spending of its subscribers.
- Police force deploying criminal detection CCTV systems at train stations.
- Financial institutions using automated risk assessment systems to approve loan applications.
- Medical practitioner using automated diagnostic systems.
- Car manufacturer deploying automated robotics assembly lines.

B.4 **End user**

An organisation or individuals that is impacted by the decisions made through the application of AI system or products.

Examples of end-users:

- Consumers and subscribers to online video platforms
- Members of public travelling through train station
- Consumer credit loan applicants
- Patients with tumour like symptoms
- Workers in a car assembly plant

Appendix C Chain of assurance compliance models

Chains of assurance vary in terms of motivation, the authority which grants a certification mark, and the rigour and liability to which compliance may be held to.

C.1 Self-declaration branding and initiative

Actor: A single company

Motivation: The motivation is usually aimed at better brand reputation towards its end consumers.

Certifying authority: A single company aiming to hold its supply chain and ecosystem partners to a standard that it has specified.

Rigor and liability: The rigour for assurance varies, but often this is in the form of a self-declaration enforced with contractual terms and conditions.

Examples: HDR10 from Samsung and the recent initiatives for privacy traffic lights by Apple for its app store.

C.2 Industry or sector-specific standards

Actor An industry consortium, consisting of many companies.

Motivation: Aims to enable a common standard to enable scale, and better interoperability between products and suppliers.

Certifying authority: The certification mark grant authority in this case is usually an evaluation body appointed by an industry consortium. While the process is a voluntary one, there is significant market friction involved if the said standard is generally recognised by consumers.

Rigor and liability: Assurance usually consists of compliance tests suites as well as interoperability tests through independent test labs.

Examples: HDMI and HDCP.

C.3 Consumer protection or other legislation with legal mandate

Actor: Government

Motivation: Where governments feel the need to impose baseline safety standards, usually for consumer protection,

product liability and compliance guidelines which are enshrined in legislation.

Certifying authority: Government

Rigor and liability: Failure to comply can result in lawsuits and penalties as outlined in legislation.

Examples: CE compliance regime for electronic goods.

Notes

[1] The scope of this white paper is limited to current AI development practices. The paper does not cover more speculative ethical questions regarding AI, such as superintelligence or the singularity.

[2] **The Arm AI Trust Manifesto**, published 2019.

[3] **'World Economic Forum launches new global initiative to advance the promise of responsible artificial intelligence**', published 2021.

[4] **IBM Research Paper Feb 2019, FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity**

5 https://arxiv.org/pdf/2004.07213.pdf

[6] **Ethics Guidelines for trustworthy AI**, published 2019.

[7]arXiv:2007.08745v3 [cs.CR] 26 Oct 2020.

[8]arxiv:1807.09173.

[9]arXiv:1707.08945

[10]arXiv:1412.6572

[11] arxiv:1805.02628

[12] **The Arm AI Trust Manifesto**, published 2019.

[13] **https://jair.org/index.php/jair/article/view/12202/26642**

[14] arxiv:2003.00898

[15]arXiv:1904.07204v

16 Barth-Jones, Daniel. The 're-identification' of Governor William Weld's medical information: a critical re-examination of health data identification risks and privacy protections, then and now. 2012.

[17] Machanavajjhala, Ashwin, and Kifer, Daniel, and Gerhrke Johannes. L-Diversity: privacy beyond k-anonymity. 2007.

[18] Samarati, Pierangela, and Sweeney, Latanya. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. 1998.

[19] Li, Ninghui, and Li, Tiancheng, and Venkatasubramanian, Suresh. T-Closeness: privacy beyond k-anonymity and l-diversity. 2007.

[20] Machanavajjhala, Ashwin, and Kifer, Daniel, and Gerhrke Johannes. L-Diversity: privacy beyond k-anonymity. 2007.

[21] Dwork, Cynthia, and McSherry, Frank, and Nissim, Kobbi, and Smith, Adam. Calibrating noise to sensitivity in private data analysis. 2006.

[22] Computational indistinguishability. See: **https://en.wikipedia.org/wiki/Computational_indistinguishability** or any standard textbook on the theory of cryptography.

[23]Apple. Differential privacy overview. See: **https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf**. Accessed 6th April 2021.

[24] Ding, Bolin, and Kulkarni, Jana, and Yekhanin, Sergey. Collecting telemetry data privately. 2017.

[25] Hawes, Michael B. Implementing differential privacy: seven lessons from the 2020 United States census. 2020.

[26] McMahan, H. Brendan, and Moore, Eider, and Ramage, Daniel, and Hampson, Seth, and Agüera y Arcas, Blaise. Communication-efficient learning of deep networks from decentralized data. 2016.

[27] Yao, Andrew Chi-Chi. Protocols for secure computations. 1982.

[28] See, for example, Gascón, Adrià, and Schoppmann, Phillipp, and Balle Borja, and Raykova, Mariana, and Doerner Jack, and Zahur, Samee, and Evans, David. Privacy-Preserving Distributed Linear Regression on High-Dimensional Data. 2016.

[29] Guestrin Carlos, Singh Sameer, Ribeiro Marco Tuli, Local Interpretable Model-Agnostic Explanations, https://arxiv.org/abs/1602.04938

[30] Miller Tim, Contrastive Explanation: A Structural-Model Approach, **https://arxiv.org/abs/1811.03163**

[31] Homma, Toshiteru, and Atlas, Les, and Marks II, Robert. An artificial neural network for spatio-temporal bipolar patterns: application to phoneme classification. 1988.

[32] Hochreiter, Sepp, and Schmidhuber, Juergen. Long short-term memory. 1997.

[33] Lynch, Clifford A. When documents deceive: trust and provenance as new factors for information retrieval in a tangled web. 2001.

[34] Cheney, James, and Chiticariu, Laura, and Tan, Wang-Chiew. Provenance in databases: why, how, and where. 2009.

[35] Verifiable Data Audit **https://deepmind.com/blog/article/trust-confidence-verifiable-data-audit**

[36] Arm TrustZone.

[37] Intel Software Guard Extensions. See: . Accessed 6th April 2021.

[38] AMD Secure Encrypted Virtualization. See: **https://developer.amd.com/sev/**. Accessed 6th April 2021.

[39] Amazon AWS Nitro Enclaves. See: **https://aws.amazon.com/ec2/nitro/nitro-enclaves/**. Accessed 6th April 2021.

[40] Cosmian. See: **https://cosmian.com**. Accessed 6th April 2021.

[41] Decentriq. See: **https://decentriq.com**. Accessed 6th April 2021.

[42] IOTEX. See: . Accessed 6th April 2021.

[43] Scalys. See: **https://scalys.com**. Accessed 6th April 2021.

[44] SCONE. See: **https://scontain.com**. Accessed 6th April 2021.

[45] Microsoft Azure confidential computing. See: **https://azure.microsoft.com/en-gb/solutions/confidential-compute/**. Accessed 6th April 2021.

[46] Amazon AWS Nitro Enclaves. See: . Accessed 6th April 2021.

[47] Sen, Youren, and Tian, Hongliang, and Chen, Yu, and Chen, Kang, and Wang, Runji, and Xu, Yi, and Xia, Yubin, and Yan, Shoumeng. Occlum: secure and efficient multitasking inside a single enclave of Intel SGX. 2020.

[48] Silicon Angle. Google debuts Confidential VMs that keep data encrypted while it's in use. See: . Accessed 6th April 2021.

[49] **https://community.arm.com/developer/ip-products/processors/b/ml-ip-blog/posts/using-psa-security-toolbox-to-protect-ml-on-the-edge**

[50] **https://www.arm.com/blogs/blueprint/metaverse**

[51] **Ethics Guidelines for trustworthy AI**, published 2019.

[52] **National Artificial Intelligence Act of 2020**, 116th Congress (2019-2020).

[53] **https://www.nist.gov/itl/ai-risk-management-framework**

[54] *The New Generational AI Development Plan* published by China's State Council, July 8, 2017.

[55] *The Development of Responsible AI: A New Generation of AI Governance Principles* published by the Ministry of Science and Technology (MoST), June 17, 2019.

[56] On March 28, the MoST held the first meeting of the New Generation AI Governance Expert Committee, chaired by Lan XUE, Dean of Schwarzman College, Tsinghua University.

[57] *The Cybersecurity Standard Practice Guide – Guidelines for Ethical Security Risk Prevention of Artificial Intelligence* released by TC260, January 5, 202.1

[58] The Beijing Academy of Artificial Intelligence is guided and supported by the MoST and Beijing municipal government.

[59] *The Beijing AI Principles* released by the BAAI, May 28, 2019.

[60] *The Joint Pledge on Artificial Intelligence Industry Self Discipline* (Draft for Comment) released by the AIIA, May 31, 2019.

[61] The Trustworthy AI Operation Guidelines (V0.5) and the Management Measures for Trusted AI Demonstration Zone, published by AIIA, August 2020.