MAGF – A Call and Proposal for Assurance Information Sharing Standards

Multi-Actor Governance Framework - Value chain communications for regulatory compliance and risk mitigation of AI systems

Working Group Contributors (Alphabetically)

Matthew Allison, Vodafone
Xin Chen, Huawei
Emanuela Giradi, PopAI
Catriona Gray, Global Digital Foundation
Rasmus Hauch, 2021.ai
John Higgins, Global Digital Foundation
Nish Imthiyaz, Vodafone
Ansgar Koene, EY
Enrico Panai, BeEthical
Stephen Pattison, ARM
Alessio Tartaro, BeEthical
Roger Whitehead, Version 1
Rob Wortham, Global Digital Foundation

Contents

About this white paper **3**
Problem Analysis **4**
Value Chain Accountability 5
Information Sharing to Improve Value Chain Accountability 6
Current and Near-future Landscape **7**
Model Cards, System Cards and Unified Frameworks 7
Human-Centred Approaches 8
Definition of Terms **9**
MAGF Dimensions **12**
The Assurance Information Sharing Process 12
Assurance Information Push and Pull 12
Typical Value Chain Configurations 13
AI Component Versioning 14
Managing Legacy Systems 15
Security 16
Assurance Information Content 16
A Step-by-Step Approach 16
The Distributed AI Value Chain 17
A Best Efforts Approach 18
Assurance Information Format 19
The Need for a Readable Format 19
Talk Like The Rest of Us 20
Example Use Case Scenarios **21**
Case study 1: Social Housing Allocation 21
Case study 2: Recruitment Decision Support System 23
Case study 3: Public Sector Citizen Chatbot 24
Case study 4: An Autonomous Vehicle Automated Parking System 25
Case study 5: Optical Character Recognition (OCR) number plate recognition 28
Recommendations **30**
Summary **30**
References **31**
Annexes **33**
Annex A - Mapping AI actors across industry, regulations and standards 33
Notes to Annex A 35

About this white paper

In this paper we motivate the need for a Multi Actor Governance Framework (MAGF). The purpose of a MAGF is to enable effective assurance across the entire AI value chain. This will enable relevant organisations to make informed, risk based decisions that meet their formal legal requirements and own business needs. Use of such a framework will also increase organisational transparency and demonstrate a responsible approach to the development and deployment of AI based solutions within the AI ecosystem, see Figure 1.

This paper is targeted at all actors in AI value chains responsible for meeting and demonstrating regulatory requirements, together with those who may not (yet) fall under any specific regulatory jurisdiction, but who nevertheless want to pursue best practice as an active risk mitigation strategy. It identifies the need for standard provisions on format, process and substantive content for information related to AI systems development, testing, certification and deployment.

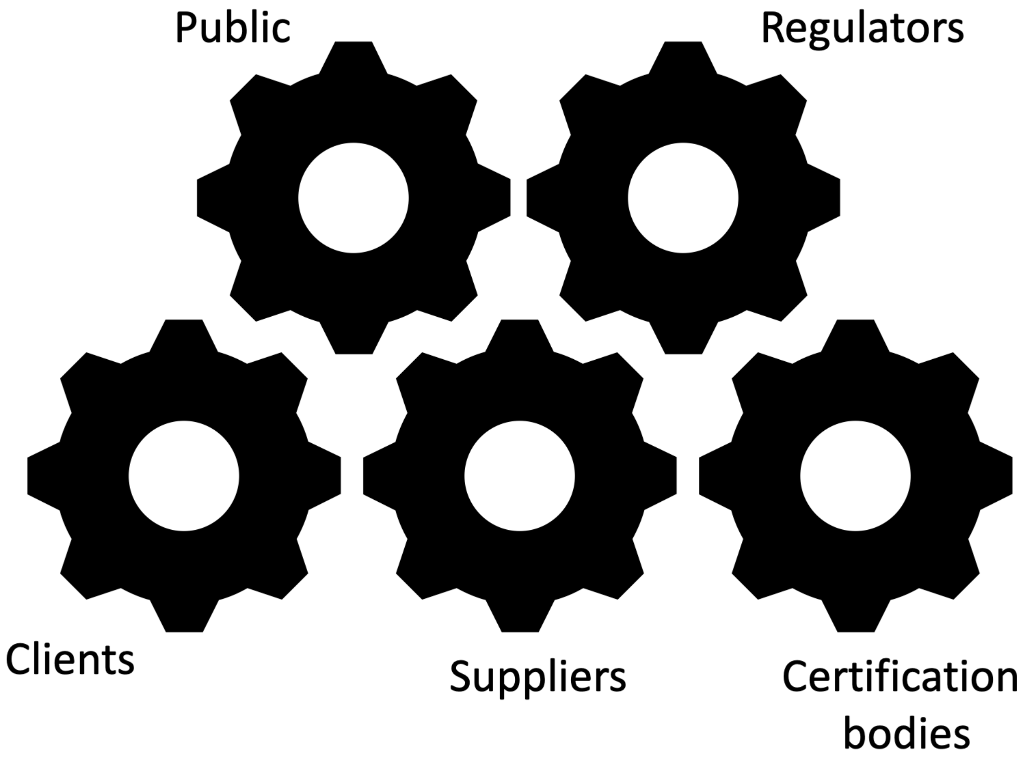


Figure 1 - The AI assurance ecosystem actors

Problem Analysis

Transparency has long been held as a fundamental principle and value to be pursued in AI governance. Policy frameworks like the UNESCO Recommendation and the NIST AI Risk Management Framework recognise the importance of ensuring appropriate information about an AI system, including its outputs, is made available to those who interact with it. Transparency of information is key to enabling accountability and building trust, see Figure 2.

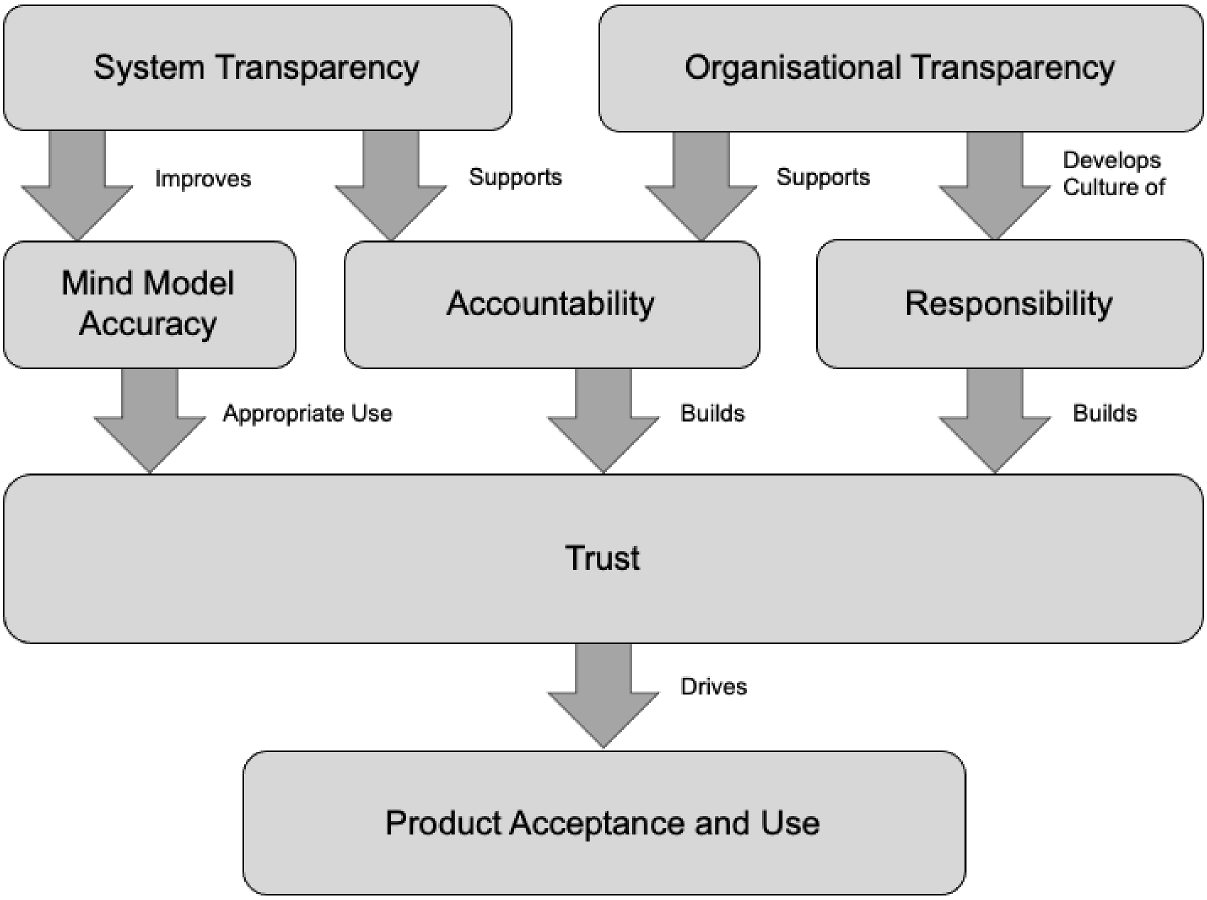


Figure 2 - The role of transparency in building trust, leading to improved product acceptance and uptake (Wortham, 2020). Reproduced with permission.

To that end, various tools, mechanisms, and frameworks have been developed to enable the sharing of information about AI systems, including information related to their input data, components and aspects of their model performance. In the following section, we outline the purpose and workings of different transparency and information sharing measures. We then explore the extent to which they can be appropriate and sufficient for tackling the problem of value chain accountability.

Value Chain Accountability

Value chain accountability has emerged as a key challenge for AI governance.[1] The production of any AI system will involve the collaboration of many actors, and AI value chains can take many forms. AI systems can be produced and distributed through many different configurations of actors and flows of information. It is relatively rare for an AI system to be built entirely ‘in-house.’ In many cases, AI systems will instead rely on an application programming interface (API), or be built according to customer problem definition or specifications. The training and retraining of AI models can also be undertaken in many different ways and with different data sources. For example, one developer might write code for an AI system without pre-training it, while another might develop an AI system using initial data provided by a customer. In other scenarios, a pre-trained model might be updated on an ongoing basis with data generated through user interaction and input.

> *Value chain accountability is an important prerequisite for making informed risk-based decisions. This is a key challenge for AI governance.*
>
> *To be able to hold others accountable, and to be accountable for their own conduct, actors will need the right information, at the right time, and in the right format.*

We understand value chain accountability as the processes and mechanisms by which all actors across AI value chains can be appropriately held to account by other actors and stakeholders. Accountability has two components: “the requirement to *give* an account, and the requirement to *maintain* an account giving relationship” (Wortham, 2020). Achieving both is an important prerequisite for making informed risk-based decisions. The variety and complexity of AI value chains can make it difficult for actors to understand and fulfil their own obligations. As a result, their ability to make informed decisions that best meet their own business needs is restricted. This is why information sharing is so key. To be able to hold others accountable, and to be accountable for their own conduct in the AI value chain, actors will need the right information, at the right time, and in the right format.

[1] See Brown, Ian (2023) ‘Allocating accountability in AI supply chains: a UK-centred regulatory perspective’ https://www.adalovelaceinstitute.org/wp-content/uploads/2023/06/Allocating-accountability-in-AI-supply-chains-June-2023.pdf

Information Sharing to Improve Value Chain Accountability

The close links between the problem of value chain accountability and practices of information sharing are identified in both academic literature and in policy documents. Cobbe et al. (2023), for example, contend that a significant challenge for governance and accountability mechanisms in AI value chains is what they call the accountability horizon i.e., “the point beyond which an actor cannot ‘see’, which depends on the actor and the chain.” As their argument goes, even if actors know with whom they are directly interacting and contracting, they may not know about those at earlier or later stages in the value chain. Actors operate with incomplete information about the AI products and services they are involved in the production and distribution of. This makes the task of comprehensive and reliable risk assessment challenging.

Downstream actors responsible for deployment often lack access to the models they depend on. There are few incentives to make information about different aspects of performance available (see section below Model Cards, System Cards and Unified Frameworks). Recent analysis[2] by Future of Life Institute examined the Terms of Service of major general purpose (foundation model) developers and found that they fail to provide downstream deployers with appropriate assurances about the quality, reliability, and accuracy of their products or services.

As Cobbe et al. (2023) go on to argue, the mechanisms needed to record, process, and provide information about AI systems are often highly contextual. Upstream actors may lack information about downstream use cases and application contexts needed to anticipate possible harms.

Recent work by Hacker et al. (2023) identifies the same set of problems in relation to the AI Act:

> *“[I]ndividual actors in the AI value chain may simply not have the all-encompassing knowledge and control that would be required if they were the sole addressees of regulatory duties [93]. This more abstract observation also shows that shared and overlapping responsibilities may be needed […] the only way forward are collaborations between LGAIM [(large generative AI model)] providers, deployers and users with respect to the fulfilment of regulatory duties.”*

[2] See Future of Life Institute, October 2023, Can we rely on information sharing? https://futureoflife.org/ai-policy/can-we-rely-on-information-sharing/

Requirements for information sharing feature in the political compromise agreement of the AI Act reached in December 2023. All so-called general purpose AI systems (i.e., foundation models) must meet certain transparency requirements, such as drawing up technical documentation, complying with EU copyright law and disseminating detailed summaries about the content used for training. The more powerful models (i.e., those posing "systemic risks") must meet even stricter testing and information sharing requirements.

It is clear, then, that scholars, practitioners, and policymakers recognize the need for information sharing mechanisms as a key enabler for value chain accountability and regulatory effectiveness.

Current and Near-future Landscape

AI documentation tools can take many forms and take different elements as their main focus e.g., models, methods, data or systems as a whole.

Model Cards, System Cards and Unified Frameworks

The concept of model cards was first proposed by academic researchers at Google[3] as a means of allowing developers of machine learning models to clarify their intended use cases and minimise their usage in contexts for which they are not well suited. Essentially, model cards are a form of documentation that can accompany a model. As well as disclosing the context in which models are intended to be used, they give details of relevant information including performance evaluation procedures. While the original formulation of model cards was focused on performance metrics related to bias and fairness, other metrics can be included.

A similar approach can be found in the publication of system cards, originally conceived by Meta. Their prototype system card was designed to explain instagram feed ranking, and they now have 22 system cards in total to help users better understand AI's role in many Instagram and Facebook features. OpenAI also make use of their own version of system cards[4]. The term has also been applied to describe a proposed unified framework for formal audits of AI based decision-aiding systems.

[3] Mitchell M. et al. (2019) FAT* '19: Conference on Fairness, Accountability, and Transparency, January 29--31, 2019, Atlanta, GA, USA https://doi.org/10.1145/3287560.3287596.
[4] Gursoy, F. and Kakadiaris, I. A. (2022) System Cards for AI-Based Automated Decision Systems. https://arxiv.org/pdf/2203.04754.pdf.

Many of these initiatives were developed to complement existing tools such as datasheets for datasets[5] and data statements for Natural Language Processing (NLP)[6] - both devised as ways of enabling documentation of a dataset's purpose, composition, collection process, and recommended uses. Other AI documentation initiatives and proposals include dataset nutrition labels[7] and supplier declarations of conformity via AI FactSheets.[8]

Human-Centred Approaches

Richards et al.[9] (IBM) propose a human-centred methodology for creating AI Factsheets. Their approach supports documentation for AI services in addition to individual models. There are also examples of similar approaches informing practice. IBM's AI Factsheets, for example, offers customers a way of tracking details of models across the lifecycle, with metadata including the purpose and criticality of the model; measured characteristics of the data set, model, or service; and lineage of events and actions taken when the model or service is created and deployed.

Some efforts have also been made to develop system-wide frameworks for AI documentation. One such example led by Partnership on AI (PAI), the ABOUT AI project, brings together stakeholders to develop comprehensive documentation tools for AI systems. The eventual stated aim of this project was to set new industry norms for documentation in AI/ML lifecycles. However, these norms have not been established or promoted.

> *Existing transparency and information sharing initiatives offer limited help to resolve information problems presented by complex AI value chains.*

[5] Gebru, T. et al. (2018) Datasheets for Datasets. CoRR abs/1803.09010 (2018). http://arxiv.org/abs/1803.09010.
[6] Bender, E. and Friedman, B. (2018) Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. Transactions of the Association for Computational Linguistics, 6:587–604 https://aclanthology.org/Q18-1041.pdf.
[7] Holland, S. et al. (2018) The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. CoRR abs/1805.03677 (2018). http://arxiv.org/abs/1805.03677.
[8] Hind, M. et al. (2018) Increasing Trust in AI Services through Supplier's Declarations of Conformity. CoRR abs/1808.07261.
[9] Richards, J. et al. (2021) A Human-Centered Methodology for Creating AI FactSheets https://research.ibm.com/publications/a-human-centered-methodology-for-creating-ai-factsheets

> *Very little attention has been paid to this important issue either in the EU, or more widely by regulators and standards bodies.*
>
> *A multi-actor governance framework will help actors to understand their own information requirements and obligations.*

While existing transparency and information sharing initiatives, such as the valuable work being undertaken by the Partnership on AI (PAI) are undoubtedly useful, their application for helping to resolve information problems presented by complex AI value chains is more limited. Indeed, very little attention has been paid to this important issue either in the EU, or more widely by regulators and standards bodies. The *multi-actor* governance framework (MAGF) for information sharing will help actors to understand their own information requirements and obligations.

Definition of Terms

For the purposes of this paper, we have defined important terms used in the table below. For further artificial intelligence concepts and terminology and standard AI definitions, please refer to ISO/IEC 22989,. Annex A also provides a mapping of AI actors across industry, regulations and standards.

AI value chain	A value chain is the series of stages through which an AI system accrues value. These stages may for example include: training data acquisition, data manipulation, model selection, model training, testing and validation, integration of trained model(s) within an application, deployment on a suitable platform, application configuration, application use by end user(s). A single organisational entity may be responsible for all steps involved in developing and deploying an AI system, but typically many separate entities are involved.
AI Assurance	AI assurance is the wide set of activities for ensuring that AI systems operate as intended, meet predefined quality standards, and adhere to ethical principles while minimising risks to individuals and society. AI Assurance typically comprises a set of processes and mechanisms designed to ensure the reliability, safety, and ethical

	use of artificial intelligence (AI) systems throughout their lifecycle. It involves assessing, monitoring, and managing the risks associated with AI applications to meet established standards and regulatory requirements. Importantly, AI Assurance is an ongoing process that extends beyond the initial development phase, ensuring that AI systems remain reliable, safe, and ethical in their operation, addressing new risks as they emerge and adhering to evolving regulatory requirements. AI Assurance can be distinguished from conformity (see below). It pursues a broader aim of ensuring AI systems are designed, developed, deployed, and operated with a strong emphasis on quality, ethics, and risk management. It encompasses not just meeting predefined standards but also minimising risks and ensuring that AI systems are responsible and safe throughout their lifecycle. Crucially, it underpins a more *horizontal* form of accountability between value chain actors, rather than *vertical* accountability towards regulatory authorities. Examples of Assurance Mechanisms: Testing and Verification: AI systems undergo rigorous testing to validate their performance. This can include unit testing, integration testing, and end-to-end testing to identify and address functional and non-functional issues. Ethical Impact Assessment: Assessing the potential ethical implications of AI systems, including fairness, bias, transparency, and accountability. Methods like fairness audits and algorithmic impact assessments are used to mitigate biases and discrimination. Security Audits: Evaluating AI systems for vulnerabilities and potential security risks to prevent data breaches, cyberattacks, and misuse. Explainability and Interpretability: Ensuring AI systems provide explanations for their decisions. Methods such as interpretable AI models and explainable AI techniques are used to make AI systems more transparent and understandable. Data Quality and Privacy Assessments: Verifying the quality of the training data, and implementing privacy safeguards to protect

	sensitive information, complying with data protection regulations (e.g., GDPR). Compliance Checks: Ensuring that AI systems adhere to legal and regulatory requirements specific to the industry or region in which they operate. Monitoring and Feedback Loops: Continuous monitoring of AI system performance in real-world scenarios and feedback mechanisms to adapt and improve the system over time.
AI Conformity	Conformity is about adherence to specifications such as those placed on certain AI systems in the forthcoming EU AI Act. While conformity to specific requirements is something that an actor may seek assurance about, this is only one aspect of AI assurance.
Risk Assessment and Mitigation	Risk assessment is a core component of AI assurance, and it involves: Identification of Risks: Identifying potential risks and issues associated with AI systems, such as biases, security vulnerabilities, or ethical concerns. Risk Quantification: Evaluating and quantifying the impact and likelihood of these risks. This helps prioritise and allocate resources for mitigation. Risk Mitigation: Implementing measures to mitigate identified risks, such as using bias mitigation techniques, enhancing cybersecurity measures, or deploying safeguards for privacy protection. Continuous Monitoring: Ongoing monitoring and evaluation of AI systems in real-world scenarios to detect new risks and adapt to changing conditions.
Technical standards	Technical standards are established guidelines, specifications, or benchmarks that provide a common framework for the design, development, operation, and interoperability of products, systems, or processes within a particular industry or domain. These standards ensure that various components or systems can work together seamlessly, share data, and maintain quality and safety, ultimately enhancing compatibility and reliability. In the context of software and AI systems, technical standards play a crucial role in promoting consistency, quality, and interoperability.

	Examples include IS0 9001 the international standard for a quality management system widely used during software development and ISO 27001 providing a framework for implementing an information security management system (ISMS). An ISMS critical for ensuring that software and AI systems handle sensitive data securely. Specific upcoming AI standards include ISO 42001 which specifies requirements and gives guidance on establishing, implementing, maintaining and continually improving an AI management system. Additionally, CEN/CENELEC have been requested by the European Commission to produce a set of AI standards defining conformance with the regulations in the EU AI Act.
AI component	A generic term used in this document to mean any AI related component produced by an actor. This includes AI products and services, but also AI models, AI algorithms, datasets to be used for model training/validation/testing, or anything else to which AI assurance information could usefully be attached.

MAGF Dimensions

In this section we cover the process of information sharing, the content of the shared information and the formatting and transmission method for assurance data passed between actors.

The Assurance Information Sharing Process

Firstly, we consider the process of sharing information through the AI value chain.

Assurance Information Push and Pull

When considering information sharing between actors, one of the first questions is who will initiate the transfer of information i.e. is the originator of the information responsible for pushing it out to interested parties, or are those who require the information responsible for requesting it? There are likely to be several triggers for this information flow It seems sensible that when an actor first decides to use an AI component, or contracts to use a service, then they can also request access to the assurance information; an information *pull*. However, in the case of an update to an AI component, for example a new version of a trained model, or an update to the assurance information relating to a dataset, then the actor responsible for that component should trigger an information *push* to all actors who use that component.

In each case, there is a need for the AI component to behave consistently and predictably. This is illustrated graphically in Figure 3 below.

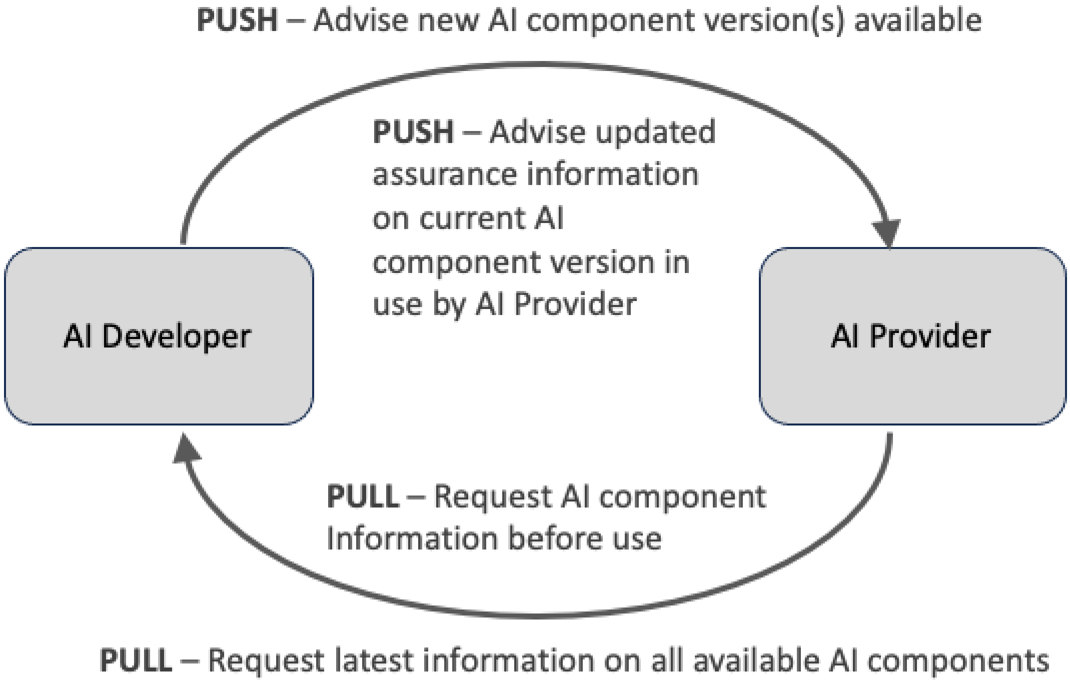

PUSH – Advise new AI component version(s) available
PUSH – Advise updated assurance information on current AI component version in use by AI Provider
AI Developer
AI Provider
PULL – Request AI component Information before use
PULL – Request latest information on all available AI components

Figure 3 - Typical push and pull events for assurance information

One complexity is that to make effective push events, the upstream AI component provider must maintain a list of all downstream consumers, i.e. all those that use their AI component. This of course would be the case where the AI component is supplied under a commercial contract, but might be much more difficult in open source scenarios. We see that some kind of centralised registry would be extremely useful here. We return to this issue in the following sections.

Typical Value Chain Configurations

There are many possible value chain configurations - though the AI Act does not expressly recognise this. A straightforward linear value chain is shown in Figure 4 below. Please refer to Annex A for definitions. In the configuration below, there may, or may not be an AI Service Provider, hence the dashed lines. Note that each of the actors may be a separate organisation, or may be different departments within one organisation, or some mix of the two.

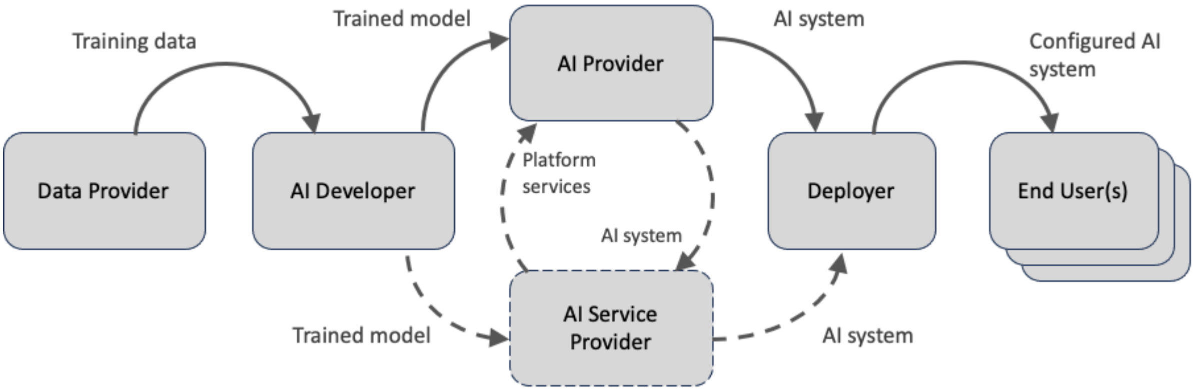


Figure 4 - The classical linear AI value chain configuration. Refer to Annex A for actor definitions.

Value chains may become considerably more complex, however, as recent work including that of Engler and Renda has outlined[10]. The AI Developer may be combining data from multiple providers, and linking the data in some way before training. Indeed, data providers may also act as aggregators of data from multiple third-party sources. The AI Provider may also produce an AI system by integrating models from multiple AI Developers, or such integration may occur within the AI Service provider. As Engler and Renda note, in these more complex scenarios, no single entity will be completely capable of evaluating and altering an AI system in order to meet the regulatory obligations of a single PHRAIS (Provider of High-Risk AI System) as required by the AI Act. It is therefore essential that information sharing occurs between actors to enable a full evaluation of regulatory obligations to be carried out.

AI Component Versioning

It is important that the assurance information related to an AI component is accurate, and relates to the correct version of that component. Datasets and definitions for trained AI models arrive in a wide variety of formats, for example Pickle, JSON, HDF5, generally without version information embedded in the content. Accurately identifying the correct assurance information therefore becomes non trivial. One

[10] Alex C. Engler and Andrea Renda (2022), Reconciling The AI Value Chain With The EU'S Artificial Intelligence Act, CEPS, Brussels https://cdn.ceps.eu/wp-content/uploads/2022/09/CEPS-In-depth-analysis-2022-03_Reconciling-the-AI-Value-Chain-with-the-EU-Artificial-Intelligence-Act.pdf

approach is to create a hash[11] from the component, and then use this hash as a unique ***assurance key*** to identify the assurance information. However, this approach would not work where frequent minor updates (tuning) of an AI component takes place, without any change in the assurance information. To support this model, the assurance information must be associated with many assurance keys.

> *Datasets and definitions for trained AI models arrive in a wide variety of formats, generally without version information.*
>
> *A unique hash of content provides an* ***assurance key*** *to correctly identify the assurance information.*

Some AI models are continually updated as the system operates, for example using online reinforcement learning based on user feedback. Some strategy must be developed to deal with the continually updating assurance information that relates to such a system. For example, the date range over which the system has been trained will change every day. Some means to create dynamic assurance information, rather than a static file based mechanism must be adopted in this case. This is a topic for further work.

Managing Legacy Systems

Generically, a legacy system is any outdated computing system, hardware or software that is still in use. Specifically applied to AI, this may occur when an AI system is still deployed and in use, but where one or more AI components used to build the system, particularly datasets or partially pre-trained models are no longer available. This could occur for a variety of reasons, for example due to liquidation of the supplier, data loss, cyber attack and so on. In this case the assurance information for a live system may not be dynamically recoverable from all actors in the AI value chain. Therefore it is important that this information is properly archived. One way to do this might be to use a trusted third party in an agreement similar to current software escrow agreements[12]. In the EU, All high-risk AI systems must be registered[13] within the EU in a database controlled by the EU Commission. Perhaps

[11] Hash algorithms operate by creating a fixed length value (typically an alphanumeric string) generated from a much larger input datastream. Even a small change to the datastream creates a very large change in the hash value. If the hash values are suitably long then the possibility of two datastreams creating the same hash becomes negligible, so the hash can be treated as a unique identifier (or key) for the original datastream.

[12] https://softwareresilience.nccgroup.com/what-is-software-escrow/

[13] EU AI Act, Recital 69.

this database could be usefully enhanced to maintain a repository of AI component assurance information, indexed by actor, product name and also assurance key. This is a topic for further exploration.

Security

The security requirements for assurance information flow are complex. We should assume that not all information can be made publicly accessible, therefore a permissions system must be implemented to grant appropriate access for downstream actors in the value chain, together with independent certification bodies, assurance third parties, regulators and so on. If a centralised assurance information repository is to be established, then this becomes a significant issue to resolve. The problem becomes more straightforward if information is shared directly along the value chain based on contractual arrangements. Platform based general purpose AI providers may well make their assurance information public, but we certainly cannot assume that will be the case for niche SMEs or highly competitive vertical market providers.

What we can say is that information flows must be encrypted to avoid interception, man-in-the-middle attacks and so on. Authentication of information sources must also be established, to ensure that assurance information is authentic. Both of these requirements are very common and widely solved in web applications today by use of the HTTPS web protocol in conjunction with site based digital certificates. Therefore these industry standard mechanisms are recommended.

Assurance Information Content

We now turn to a more detailed analysis of the information required to be shared.

A Step-by-Step Approach

Building on previous work published by Global Digital Foundation[14], a simple step-by-step approach is theoretically possible to derive the minimum set of assurance information content directly from legal clauses and technical tests required for regulatory compliance. For example, Article 9 of the proposed EU AI Act requires identification and assessment of risks and details of appropriate risk management steps taken to mitigate risks. This level of documentation is of course at a fairly high and possibly rather abstract level. However it can be subsequently refined to a much finer granularity once the related CEN/CENELEC standards are

[14] Higgins, J. and McDonnell, P. (2020), ‘Towards Multi-Actor AI Governance In Five Practical Steps’ and Moës, N. (2021), ‘Feedback Report - September 2021 “Multi-Actors Governance Framework” (MAGF)’

completed. These standards have already been requested by the EU Commission[15]. We can therefore envisage a very substantial, but essentially mechanical, decomposition task to dissect the detailed information requirements arising from a combination of these standards and their parent, the finalised AI Act. One would codify these information requirements as named data fields and specify an overall schema for AI assurance information. **We strongly suggest, as further work, that this codification task is undertaken for the EU AI Act once it is finalised in the coming months.** Similar, but smaller tasks could go some way to defining schemas for the recent US Government Executive Order on AI[16] which whilst still at a high level, runs to some 110 pages of detail.

The Distributed AI Value Chain

However, some significant complexity arises when we consider that most AI systems deployed in the marketplace by an AI Provider (see Annex A for definitions) are a complex assemblage of AI components developed and made available by a variety of actors within an overall AI value chain (see Definitions above). Therefore no one actor has all the required information, nor do they necessarily know whether a complete set of compliance information can be assembled by contacting all those in the AI value chain. Therefore, what information should they make available to downstream consumers of their AI components? As we explained in the current landscape section above, some would argue for an entirely bespoke, component-by-component and consumer-by-consumer approach to answer this question. However this neither scales by product, nor by market size, and certainly does not take into account the power dynamics of the AI ecosystem.

> *We advocate a cooperative, best efforts approach from all actors to deliver the most complete overall information flow, minimising risk and maximising regulatory compliance.*

There currently exists no predefined set of information that each actor in the value chain, for example categorised by the taxonomy of Annex A, should provide. We can see though that based on that taxonomy, those that provide data sets as a basis for training, validation and test data can answer questions relating to the source, completeness, errors and other characteristics of that data, whilst those

[15] GROW.H.3, December 2022, 'Draft standardisation request to the European Standardisation Organisations in support of safe and trustworthy artificial intelligence'

[16] Office of President of USA, October 2023, 'Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence'

subsequently preparing data and using it for AI system training can naturally provide information relating to those activities, and so on.

A Best Efforts Approach

In the absence of prior definition, AI component suppliers can take a best efforts approach to complete an overall schema to the best of their ability. They can also pass information received from upstream suppliers to downstream consumers.

Ultimately, the AI Provider must assemble a complete set of information, fully populating the assurance information schema. They can identify missing data items and ask for updates from those upstream, and/or they can identify that missing information in their documentation having made best efforts to retrieve it. Upstream actors can similarly take these actions (on a more limited basis), depending on their position within the value chain.

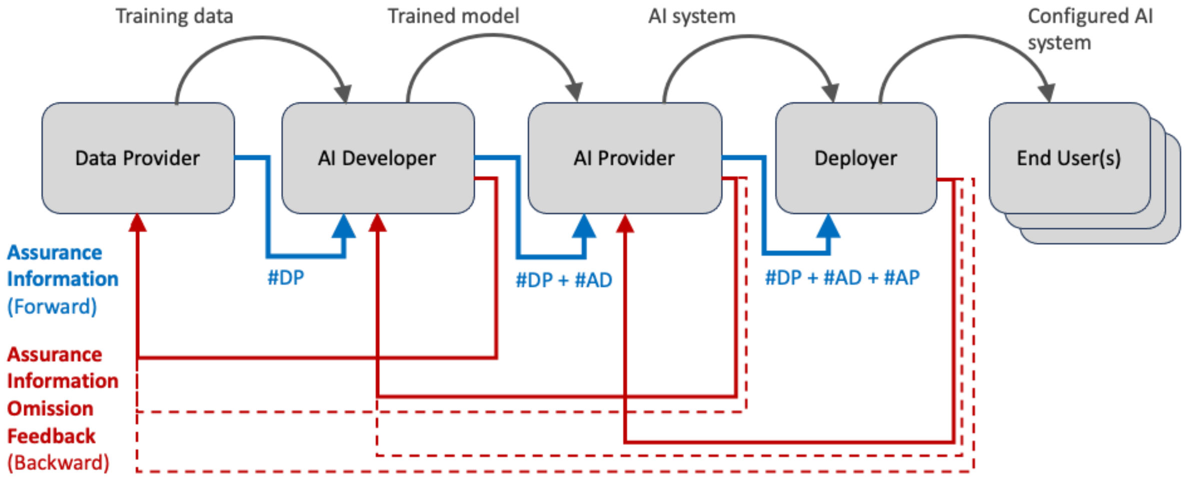


Figure 5 - Best Efforts Cumulative Completion and Omission Feedback (BECCOF).#DP indicates information from Data Provider. #AD indicates information from AI Developer. #AP indicates information from AI Provider. See Annex A for definitions.

We suggest and advocate this BECCOF approach as a cooperative, best efforts approach from all actors to deliver the most complete overall information flow, minimising risk and maximising regulatory compliance. This best efforts cumulative completion and omission reporting and feedback approach is illustrated graphically for a simple case in Figure 5 above.

Assurance Information Format

It may seem overly detailed to include a section on data formatting in a white paper. However, the uptake of inter-organisation communication standards, has much to do with the ease with which organisations can engage and participate, and the ability of the chosen format to support gradual evolution and backwards compatibility as the ecosystem evolves and matures.

The Need for a Readable Format

The rapid rise and success of text based formats such as XML and JSON demonstrates that formats that are easily readable by humans are preferred over complex APIs (Applications Programming Interfaces). It's also important to recognise that information flowing through the value chain will have a hierarchical structure with repeating groups of similarly structured information. Finally, for wide international interoperability, any text based communication method must support internationalisation (i.e. support for international character sets).

A successful inter-organisation communication standards must be:

- *Easy to use*
- *Readable by Humans*
- *Hierarchical*
- *International*

In practise the candidate format standards are XML and JSON. XML benefits from strong schema definition (using schema definitions also written in XML), but maybe unnecessarily complex and can make for difficult reading by those unfamiliar with XML syntax. JavaScript Object Notation (JSON) is a simple and lightweight text-based data format. We recommend JSON as it is easy to read for non technical users. JSON can also be supplemented by JSON Schema[17] to define allowable data elements and structures.

[17] JSON Schema is an IETF standard providing a format for what JSON data is required for a given application and how to interact with it. Applying such standards for a JSON document lets you enforce consistency and data validity across similar JSON data. Although currently still technically a draft standard, JSON Schema is widely used and well supported.
See https://json-schema.org/overview/what-is-jsonschema

Talk Like The Rest of Us

Finally, a word about how assurance messages are actually sent and received. Once again we should reach for the simplest and most widely adopted technologies for this purpose. The web-based communication protocol HTTPS is used almost universally for communication with web sites and benefits from SSL security which provides both encryption and host authentication.

Example Use Case Scenarios

We suggest five use cases to demonstrate how the framework would operate:

- A local authority social housing allocation system;
- A recruitment decision support system;
- A system using a generative transformer model to provide customer interfacing services (a public sector citizen chatbot);
- An Automated Parking System for an Autonomous Vehicle;
- Optical Character Recognition (OCR) number plate recognition.

For each case study, we first describe the scenario and then identify the actors in terms of the ontology developed in Annex A. We identify some of the key obligations of each actor, particularly with reference to the EU AI Act, as this regulation provides the most specific detail at this time. Finally we suggest how the application of a MAGF can assist in AI value chain assurance.

Case study 1: Social Housing Allocation

Scenario	A local authority (LA) in a European Union member state wants to use AI to help allocate social housing provision. Demand in the LA area greatly exceeds available housing stock. The LA is under a legal obligation to allocate provision in a targeted manner according to criteria set out in legislation. These criteria include: income, age, immigration status, and tenant's housing history. Preferences may be given to applicants who are experiencing homelessness or who need specific accommodation for health or disability reasons. The LA also tries to match applicants to the most appropriate properties e.g., families with young children would need schools nearby. The LA decides to procure from a private company (PC) an AI system that will use historical data stored in the LA's asset management system. The AI system is designed to help with prioritisation and allocation by providing real-time, dynamic information, and to help predict future demand and capacity. The software code has been written entirely by PC and the model has already been fully trained. Access to the AI system is enabled through an API and no changes to the model can be made by the LA.

Actors	In this scenario, it is important to determine which roles each actor would be fulfilling (**see Annex A: Taxonomy**). PC would likely be an AI Developer, an AI Provider and an AI Service Provider. Of these roles, only AI Provider comes with any legal obligations. According to our Taxonomy, the LA would be a Deployer (‘User’ under the Commission’s proposed AI Act text). The application would be classified as high-risk under the AI Act.
Obligations	From the perspective of the PC, as a provider of a high-risk AI system (PHRAIS) under the AI Act, it would be responsible for the majority of legal requirements. These include provisions on risk management, data governance, technical documentation, record keeping, transparency, accuracy and robustness. Providers must also go through a formal process of ex ante conformity assessment, registration, and nomination of an authorised representative. From the perspective of the LA, it must fulfil requirements as a ‘user’ (Deployer) under the AI Act. Most of these requirements are set out in Article 29, and include a requirement to use the system in accordance with the instructions provided. In addition to the AI Act, the LA would have numerous other general legal obligations, including those under administrative law, and equality and non-discrimination law. They would very likely need to comply with the Procurement Clauses for AI, for example. The LA would also have to consider reputational implications and democratic accountability.
Framework application and benefits	The framework could be used to assist the LA in its procurement of the system. It could use the framework to guide its negotiations with different suppliers, and to help draft contractual terms. From the suppliers’ perspective(s), the framework helps to reduce the risk of engaging with the LA to deliver the system. The supplier has a clear understanding of the information to be provided, and the format and timeliness of the information delivery. This perceived risk reduction may result in more favourable commercial terms for the LA.

Case study 2: Recruitment Decision Support System

Scenario	The owners of a privately held business in the UK (UK-B) want to reduce their recruitment costs. After reviewing options available, they decide to purchase a subscription to an AI-enabled platform, provided by a US-based business (US-B). The platform can be used for a full range of recruitment activities that include: ● Using generative AI to write job descriptions and advertising material; ● Candidate sourcing and matching; ● CV and application screening e.g., extracting relevant information, ranking candidates, and filtering out unqualified applications; ● Testing and assessment e.g., video interviews using facial recognition, speech analysis, or sentiment analysis. The software code has been written entirely by US-B. However, the model will continue to be updated based on data provided by UK-B and other customers. UK-B is considering whether to expand its operations into the EU market in the near future.
Actors	In this scenario, US-B would likely be an AI Developer, an AI Provider and an AI Service Provider. If this scenario were taking place in the EU, only the AI Provider would come with specific legal obligations. According to our Taxonomy, UK-B would be a Deployer (‘User’ under the AI Act). The application would be classified as high-risk under the AI Act.
Obligations	As UK-B and its operations are currently entirely UK-based, it would not be required to conform to the AI Act. However, it would have many other legal obligations under UK law. In particular, UK equality and non-discrimination law, and data protection law, would govern UK-B’s decision-making and processing of all personal data. Assuming the business is based in England, the most relevant regulators would be the EHRC and the ICO, who may have guidance for businesses on AI and recruitment. Beyond its legal obligations, UK-B would also have to consider reputational implications. When considering its entry into the EU

	market, UK-B would, as a minimum, need all information relevant to meeting the requirements under EU law, including the AI Act.
Framework application and benefits	When UK-B considers whether to expand its operations into the EU market in the near future, then as a minimum, UK-B would need all information relevant to meeting the requirements under EU law, including the AI Act. They can use the framework to request this information from US-B.

Case study 3: Public Sector Citizen Chatbot

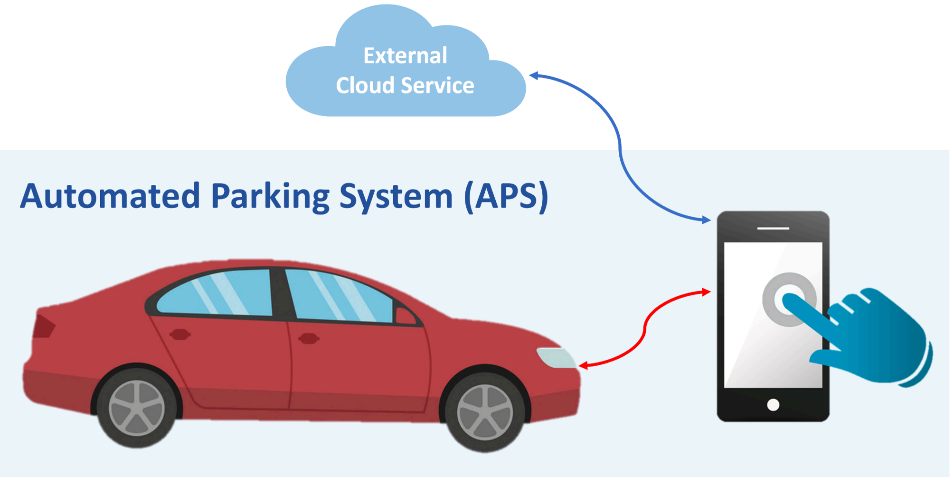
Scenario	A European town Council develops a ‘smart agent’ with a primarily speech based interface for residents, accessible from their mobile phones. It allows residents to ask questions about council services based on council publications, and also allows access to demographic and other public data relating to the local area. Finally, it allows interaction with local council run services, including registering a young child for initial entry to local schools and finding real time availability for council run car parks. To keep costs down, the Council elect to use in-house software developers, who develop using a new large language model (LLM) agent capabilities, which allow systems to be designed using natural language and by supplying the system with files containing details of Council services, including details of how to carry out the supported interactions on existing council web sites. The agent is made available without charge in the LLM provider’s app store. The Council recognises the need for robust testing, and uses a third party company (TestCo GmbH) to carry out ‘red teaming’ and similar testing using their expertise.
Actors	In this scenario, the Council would be an AI Developer, an AI Provider, a Deployer and a Data Provider, whilst the LLM provider would be an AI Service Provider. There would also be a further Data Provider - the provider of the training data for the underlying LLM foundation model, however the town Council would have little visibility of this entity. Local town residents would be End Users.

	TestCo would have no designation within the taxonomy of the AI Act. The application would be classified as high-risk under the AI Act.
Obligations	The town Council would bear the majority of the obligations under the AI Act, as they would classify as the AI Provider. Despite the LLM being a foundation model, it has been substantially adapted for this one application, making the Council the PHRAIS. It is worth noting that TestCo's only responsibilities and liability would relate to its contract with the Council, and the council would not be able to offload its responsibility with respect to testing by using a third party.
Framework application and benefits	The framework could be used to share information between the LLM provider and the town Council. In this case the Council would pull information relating to the LLM and the new agent functionality. It would then enhance this information based on the additional training data supplied by the Council, together with information relating to the external systems being interfaced. The assurance information from the Council would then be used by TestCo as a basis for designing detailed testing of the overall system. The use of the framework would reduce the unknowns in this project, therefore lowering the risks. It would also form the basis of information provided for End Users, and the auditors of the Council's affairs.

Case study 4: An Autonomous Vehicle Automated Parking System

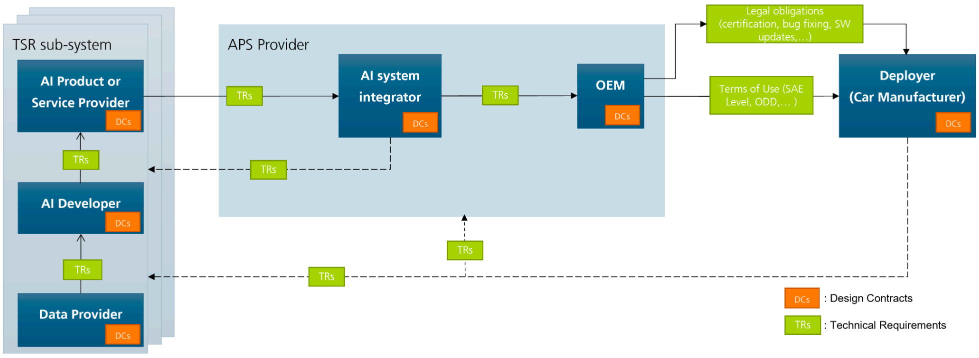
Note, this case study is based on an automated parking system (APS) described in current work being produced for future publication by the Fraunhofer Institute for Cognitive Systems (IKS) and is reproduced here with permission.

Scenario



A car manufacturer (CarCo) wants to add automated parking capabilities to their electric vehicles. To park, the user exits the vehicle, searches for free parking slots in the APS mobile phone app, selects a parking slot and activates the APS function to drive autonomously to the parking slot and park the car. To retrieve the vehicle, the user activates the APS app and indicates the pickup time and location.

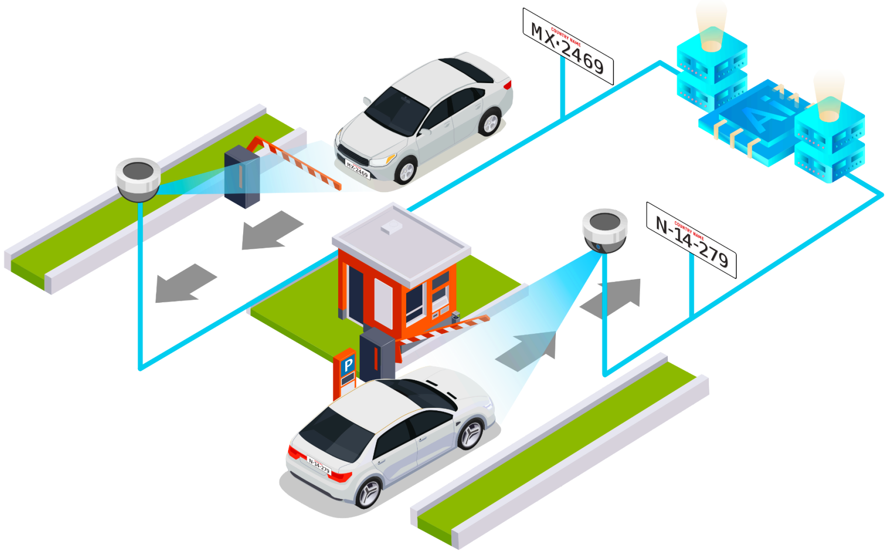
CarCo contracts with APS Provider to use their product to achieve the desired functionality. A partial overview of the supply chain with traffic sign recognition (TSR) as a subsystem for the APS Provider is shown below. Note that many subsystems may exist which have technical requirements with the APS Provider and other stakeholders. Only the TSR subsystem is shown in the figure below.


TSR sub-system
AI Product or Service Provider
DCs
TRs
AI Developer
DCs
TRs
Data Provider
DCs
APS Provider
TRs
AI system integrator
DCs
TRs
OEM
DCs
TRs
TRs
TRs
Legal obligations (certification, bug fixing, SW updates,…)
Terms of Use (SAE Level, ODD,…)
Deployer (Car Manufacturer)
DCs
DCs : Design Contracts
TRs : Technical Requirements

This system would certainly be classified as high-risk under the AI Act, due to the risks of injury to pedestrians, damage to property and so on.

Actors	APS Provider is placing an AI component on the marketplace to be deployed in the marketplace by car manufacturers including CarCo. They are thus classified as an AI Provider, and an AI Developer. CarCo is thus a Deployer. The TSR subsystem is provided by another third party who also make their product available in the marketplace, and so are also an AI Provider and AI Developer.
Obligations	Contractual and design obligations exist between all the parties as shown in the diagram, creating a complex system of obligations. In terms of the overall functionality of the car, CarCo have carried out the contracting and integration of these subsystems within the overall vehicle design, and so product liability from the consumers' (End Users') perspective lies with CarCo.
Framework application and benefits	In this application, use of the framework will enormously assist all parties in better understanding the details of these subsystems from a risk perspective. The information shared through the framework will help in finalising contractual details regarding performance and liability. The information will also be valuable for third party assessors in relation to legal obligations under the AI Act or similar regulatory regimes.

Case study 5: Optical Character Recognition (OCR) number plate recognition

Scenario	 The implementation of Optical Character Recognition (OCR) number plate recognition technology is primarily focused on private and restricted environments such as mall parking lots. Here, it automates the capture and recognition of vehicle licence plates as they enter and exit. This system is particularly suited for spaces like smart communities, modern shopping malls, and automotive service centres. The OCR-NPR is an AI-enhanced Optical Character Recognition system. This system facilitates intelligent management and supports automatic payment processing. It significantly contributes to reducing the duration of vehicle traffic, enhancing the overall flow efficiency, and can operate around the clock in unattended spaces. The automation at its core also ensures a low-cost solution, making it economically viable. In summary, an AI-enhanced OCR number plate recognition system offers a rapid and effective solution for managing parking in malls. This system significantly cuts down waiting times and enhances the overall customer experience. The AI system implies the following steps: 1. Camera Setup: Cameras are installed at the entrance and exit points of the mall parking. These cameras are equipped with AI algorithms that can recognize licence plate characters.

	2. Image capture: As a vehicle approaches the camera, the system captures an image of the licence plate. 3. Image processing: The image captured by the camera is then processed using AI algorithms to identify the licence plate number. 4. Plate comparison: The identified licence plate number is then compared to a database of registered vehicles. If the vehicle is registered, the system will allow it to enter or exit the parking lot. If it is not registered, an alert will be sent to the security team. 5. Automated Payment: In some cases, the system can also be integrated with an automated payment system, allowing customers to pay for their parking using their licence plate number as a unique identifier. 6. Data storage: The system stores data on the licence plates of all vehicles that enter and exit the parking lot, providing valuable information to mall management for traffic analysis and optimization.
Actors	Provider, importer, and distributor play a pivotal role in ensuring key aspects of AI systems, like reliability and transparency, are maintained. Establishing a mature and stable process for allocating responsibility is crucial for enhancing the trust infrastructure in the market. This approach aligns with the guidelines set out in Recital 60 of the EU AI Act, which suggests that relevant third parties should collaborate effectively with providers and users.
Obligations	There are various contractual, non-contractual and regulatory obligations that bind parties involved in creating these AI systems. These obligations form a matrix of responsibilities,and their fulfilment means ensures that every element of the system is accounted for and meets the necessary standards. This complex structure is necessary to manage the multifaceted nature of AI system development, ensuring accountability and quality in every stage.
Framework application and benefits	Utilising this framework provides significant advantages. It aids all involved parties in comprehensively understanding the subsystems from a risk perspective. This understanding is crucial in finalising contractual elements related to performance and liability. Moreover, the shared information within this framework is invaluable for third-party assessors. It supports them in evaluating compliance with legal obligations, especially under regulations like the AI Act.

Recommendations

- All actors, particularly regulators and standards bodies, should work together to analyse and deepen their understanding of information sharing problems present across AI value chains.

- Additional tools and resources should be developed and made available to help businesses to understand their own information requirements and obligations.

- Once the proposed AI Act is finalised, a systematic analysis of its provisions related documentation should be undertaken with the purpose of codifying all requirements.

- Further work should be undertaken to develop appropriate strategies to deal with the dynamic nature of assurance information requirements.

- The European Commission should consider enhancing its database of AI systems to include AI component assurance information.

- To maintain the security of flows of assurance information, we recommend the use of the HTTPS web protocol in conjunction with site based digital certificates. When it comes to determining the appropriate format for assurance information sharing, we recommend the use of JSON.

- To understand and meet their contractual liabilities, all parties will require clear and agreed definitions and agreement about what information must be shared, at which points and through which mechanisms. We recommend that standards bodies, such as ISO and CEN/CENELEC, include value chain information sharing in their work.

Summary

Organisations involved in the AI value chain need to make informed, risk based decisions that meet their formal legal (regulatory) requirements and own business needs. It is essential to recognise the importance of a transparent and timely flow of information between actors to facilitate assurance through the value chain. Very little attention has been paid to the important issue of information sharing through the value chain either in the EU, or more widely by regulators and standards bodies. A multi-actor governance framework (MAGF) will help actors to understand their own information requirements and obligations. Also, use of an AI assurance information sharing framework will increase organisational transparency and demonstrate a responsible approach to the development and deployment of AI based solutions.

In this paper, we recognise and consider the practical issues relating to complex value chain configurations, particularly noting the inherent power dynamics between the big tech providers and SME AI developers and deployers. We consider the design of a multi-actor governance framework to support assurance information flow between actors in the AI value chain. We identify the actors involved and describe their roles and deliverables based on a taxonomy derived from international standards and the definitions in the EU AI Act. We propose a methodology to ascertain the necessary information to be shared and suggest a Best Efforts Cumulative Completion and Omission Feedback (BECCOF) approach whereby all actors in the value chain cooperate to improve transparency and demonstrate regulatory compliance. Finally we use several case studies to show how the framework can be applied and deliver benefits to all actors involved in the process. Going forward, our recommendations are as above.

References

Andrade, Norberto Nuno Gomes de, and Antonella Zarra. “Artificial Intelligence Act: A Policy Prototyping Experiment: Operationalizing the Requirements for AI Systems – Part I” (2022), at http://openloop.org/reports/2022/11/Artificial_Intelligence_Act_A_Policy_Prototyping_Experiment_Operationalizing_Reqs_Part1.pdf

Cobbe, J. et al. (2023) Understanding accountability in algorithmic supply chains, https://arxiv.org/abs/2304.14749

Hacker, P. et al. (2023) Regulating ChatGPT and other Large Generative AI Models https://doi.org/10.1145/3593013.3594067

Küspert, S. et al. (2023) The value chain of general-purpose AI https://www.adalovelaceinstitute.org/blog/value-chain-general-purpose-ai/

Brown, I. (2023) Expert explainer: Allocating accountability in AI supply chains https://www.adalovelaceinstitute.org/resource/ai-supply-chains/

Portfolio of Assurance Techniques by DSIT’s Centre for Data Ethics & Innovation (CDEI) https://cdei.blog.gov.uk/2023/06/07/from-principles-to-practice-launching-the-portfolio-of-ai-assurance-techniques/

Alex C. Engler and Andrea Renda (2022), Reconciling The AI Value Chain With The EU'S Artificial Intelligence Act, CEPS, Brussels
https://cdn.ceps.eu/wp-content/uploads/2022/09/CEPS-In-depth-analysis-2022-03_Reconciling-the-AI-Value-Chain-with-the-EU-Artificial-Intelligence-Act.pdf

Wortham, R. H. (2020), *Transparency for robots and autonomous systems: fundamentals, technologies and applications*, Stevenage, England: Institution of Engineering and Technology.
https://shop.theiet.org/transparency-for-robots-and-autonomous-systems

Annexes

Annex A - Mapping AI actors across industry, regulations and standards

The table provides a mapping between the AI actors and their respective definitions, as proposed in Andrade et al's "Artificial Intelligence Act: A Policy Prototyping Experiment. Revisiting the Taxonomy of AI Actors" and the AI actors defined in other documents including:

(1) The European Commission's AI Act proposal, and the positions adopted by the European Parliament in June 2023, and the Council of the European Union in December 2022;
(2) ISO/IEC 22989 Information Technology — Artificial Intelligence — Concepts and Terminology.

It should be noted that, in certain cases, the AI Act and ISO/IEC 22989 include actors that do not have any equivalent in the Taxonomy. As well as those featured in the table below, the European Commission's proposed AI Act includes: *authorised representative* and *operator.* The first refers to a nominated party with a written mandate to carry out the responsibilities on behalf of a provider while an operator is an umbrella term covering the roles of: user, the authorised representative, the importer and the distributor.

Taxonomy	**Definition**	**AI Act**	**ISO/IEC 22989**
AI Developer	The natural or legal person that builds generic or specific AI systems at the behest of third parties at the behest of third parties or for self interest but who do not place this product on the EU market.	NA [1]	AI Producer; AI Developer [2]
AI Provider	The natural or legal person that places a generic or specific AI system on the EU market.	Provider	AI Provider [3]
		Small-scale or SME provider [4]	
AI Service Provider	The natural or legal person that provides AI support tools and/or services on demand.	NA [5]	AI Platform Provider

Taxonomy	Definition	AI Act	ISO/IEC 22989
			AI Service or Product Provider
Data Provider	The natural or legal person that provides data for training, testing and/or validating generic or specific AI systems.	NA	AI Partner [7]
			Data Provider [8]
Deployer	The natural or legal person using a specific or generic AI system to perform a particular task.	User [9]	AI Customer [10]
			AI Users
End User	The natural person operating the AI system and/or using AI system outputs to inform their actions.	User [9]	NA [11]
Subject	A natural or legal person that is directly influenced by the outcomes of an AI system.	Affected person [12]	AI Subject [13]
			Data Subject [13]
			Other Subject [13]
Importer	The natural or legal person established in the EU importing a generic or specific AI system from outside the EU and placing it on the EU market.	Importer [14]	NA
Distributor	The natural or legal person established in the EU importing a generic or specific AI system from outside the EU and making it available to a provider that places it on the EU market.	Distributor [15]	NA

Notes to Annex A

[1] In the AI Act, the provider 'develops an AI system', and in this it is similar to the AI Developer in the Taxonomy but it develops the AI system 'with a view to placing it on the market', and in this it differs from the AI Developer in the Taxonomy, who instead develops the AI systems for third parties.

[2] The AI Developer of the Taxonomy encompasses both AI producer and AI developer in ISO/IEC 22989. The two definitions concern two roles involved in the creation and distribution of AI systems. There are some differences in how they deal with distribution, market interaction, and third-party involvement.

[3] Despite the use of the same term, the definitions in the Taxonomy and in ISO/IEC 22989 differ in their scope and geographical context.

[4] The Taxonomy makes no distinction according to provider size, unlike the AI Act, which presents Small-scale providers (European Commission proposal and European Parliament amendments) and SMEs (Council of the European Union General Approach).

[5] In the AI Act, the provider always places an AI system on the market or in service under its own name or trademark. There are therefore no intermediate actors such as the AI service provider described in the Taxonomy.

[6] ISO/IEC 22989 breaks down the role of AI service providers into AI platform provider and AI service or product provider. The definitions of AI platform provider and AI service or product provider taken together broadly correspond to that of AI service providers in the Taxonomy.

[7] In ISO/IEC 22989, AI partner is a very broad category, which includes not only the Data Provider, but also AI Actors not mentioned in the Taxonomy, such as AI evaluator and AI auditor.

[8]In this case, the Taxonomy definition is more detailed than that in ISO/IEC 22989, particularly in respect of the purpose of the data provided, and the scope of the AI involvement.

[9] In the AI Act no distinction is made between deployer and end user as in the Taxonomy. The notion of user captures both of them.

[10] AI customer in ISO/IEC 22989 is a particular type of User. The definition of AI customer provides more detail on the level of interaction between a User and an AI system. An AI customer is an entity that uses an AI product or service either directly or indirectly (by providing it to other AI users). This suggests a range of possible interactions, from directly using an AI product or service to act as a middleman supplying AI systems to end users. On the contrary, the Taxonomy definition implies direct interaction with the AI system and a more hands-on use of the technology.

[11] While ISO/IEC 22989 provides several categories of Users, none of them refer to natural persons who use AI systems to inform their actions, and thus do not capture the notion of End users proposed by the taxonomy.

[12] These two definitions refer to individuals or groups that interact with or are impacted by AI systems, but they differ in their scope, whether they address a group or single person/entity, and in whether the subject is directly or indirectly impacted.

[13] ISO/IEC 22989 breaks down the concept of AI subject into various specific categories like “Data subject” and “Other subject”, providing examples of each. This categorization emphasises different ways an entity can be impacted by an AI system - either through the use of their data in training the AI or by interacting with AI-enabled products or services. On the other hand, the Taxonomy definition provides a broad and straightforward explanation without going into specifics about different types of subjects or ways they might be influenced. In addition, AI subjects in ISO/IEC 22989 are “impacted”, a term that is somewhat vague and could refer to a wide range of potential effects. The Taxonomy uses the term “directly influenced by the outcomes of an AI system”, which might suggest a more immediate or clear effect stemming from the AI system's outputs.

[14] The AI Act specifies that the importer places on the market or puts into service in Europe an imported AI system bearing the name or trademark of a natural or legal person established outside the Union. This clarification is missing in the Taxonomy.

[15] Although they employ the same term, these two definitions represent slightly different roles in the supply chain for AI systems in the European Union, and there are two main differences between them: the role in the supply chain; and the source of AI systems.