FORHUMANITY
980 Broadway #506 Thornwood, NY 10594
(+1) 9146028663

ryan@forhumanity.

FORHUMANITY EUROPE
12 rue Frederic Petit
80000 Amines, France

FORHUMANITY

# CORE AAA Governance Provider-only

CERTIFICATION SCHEME V1.0
Artificial Intelligence, Algorithmic and Autonomous (AAA) Systems

# Introduction

ForHumanity (https://forhumanity.center/) is a 501(c)(3) non profit organization and ForHumanity Europe is a French 1901 Association, dedicated to addressing risks associated with Ethics, Bias, Privacy, Trust, and Cybersecurity in Artificial Intelligence, Algorithmic, and Autonomous (AAA) Systems. ForHumanity uses an open and transparent process that draws from a pool of over 2600+ international contributors to construct audit criteria, certification schemes, and educational programs for legal and compliance professionals, educators, auditors, designers, developers, and legislators to mitigate bias, enhance ethics, protect privacy, build trust, improve cybersecurity, and drive accountability & transparency in AAA Systems. ForHumanity works to make AAA Systems safe for all people and makes itself available to support government agencies and instrumentalities to manage risk associated with AAA Systems. Our mission is to *examine and analyze downside risk associated with the ubiquitous advance of AI, algorithmic and autonomous systems and where possible to engage in risk mitigation to maximize the benefits of these systems… ForHumanity*

## Infrastructure of Trust

ForHumanity supports an infrastructure of trust predicated on the 50+ year track record of financial accounting and reporting. This infrastructure of trust is founded on a principle of jurisdictional sensitivity, which means that each sovereign nation-state or region has the right to establish their own laws, regulations, guidelines, and shared moral framework. ForHumanity affirms that right by ensuring that our certification program upholds local laws and seeks approval, where applicable, from local authorities. Key elements of Independent Audit of AI Systems are critical to ensure that it functions properly across multiple different jurisdictions, these are non-negotiable elements of the shared moral framework that constitutes Independent Audit of AI Systems and they include concepts such as transparency, disclosure, independence, risk management, and ethical oversight.

ForHumanity believes that a binary (compliant/non-compliant) set of criteria, either adopted by common practice in the marketplace or approved by the sufficient governmental authorities, and subsequently assured for compliance independently by certifying bodies (auditors), can create an infrastructure of trust for the public that assures compliance with laws, regulations, guidelines, standards, and best practices in a proactive manner when combined with the requirement for regular, mandatory, independent audits.

An infrastructure of trust, as it relates to certification, is an unconflicted process deploying a segregation of duties, conducted by certified and trained experts, that establishes a robust ecosystem that engenders trust for all citizens and protects those who have no power or control.

The infrastructure of Trust that For Humanity supports is grounded on four core tenets:

1. ForHumanity produces accessible, binary (compliant / not compliant) certification criteria that transparently and inclusively aligns laws, regulations, standards, guidance and best practice that embeds compliance and performance into practice, and is considerate of corporate wisdom, but impervious to corporate dilution and undue influence, while being mindful of the regulatory burden and dedicated to maximizing risk mitigations to humans.

2. Individuals are trained and accredited on certification criteria as experts by ForHumanity. They perform pre-audit and audit services on behalf of certification bodies and are individually held to a high standard of behavior and professionalism as described in the ForHumanity Code of Ethics and Professional Conduct - they are ForHumanity Certified Auditors (FHCAs)

3. Certification Bodies employ FHCAs to independently assure compliance with certification criteria on behalf of the public. They are licensed, independent, robust organizations that take on the task and risk, on behalf of the public, to ascertain assurance of compliance. They are held to standards of independence and anti-collusion and are further subject to third-party oversight ("watching the watchers"), by entities such as national accreditation bodies (e.g. COFRAC, UKAS, DaKKE) and ForHumanity.

4. Corporations and public sector Providers and Deployers of AAA Systems can use the criteria to operationalise governance, oversight, and accountability that helps them to achieve required conformity under the law. Compliance with ForHumanity certification schemes will create leverageable governance, oversight, and accountability that will simultaneously lead to more sustainable profitability and reduce the risk of negative outcomes for their stakeholders.

See Appendix A for more details on Roles and Responsibilities in an Infrastructure of Trust[1].

[1] Infrastructure of Trust for AI – Guide to Entity Roles and Responsibilities

# ForHumanity's Role in an Infrastructure of Trust

Founded in 2016, ForHumanity first wrote about Independent Audit of AI Systems in 2017 and it has been our primary focus since that time. We advocate for mandatory independent audits and the establishment of the aforementioned infrastructure of trust similar to those required in financial accounts and reporting.

Transforming an audit ecosystem from financial audits to process audits for AAA Systems requires thoughtful adaptation. Transformation occurs by accomplishing the following tasks:

1. Understanding how financial audit rules & standards mitigate risk, provide clarity, and translate opaque controls and processes into public trust and valuable cross-sectional comparability through third-party independent assurance
2. Understanding the risks of AAA Systems and developing rules & standards to treat and mitigate risks to stakeholders, including individuals
3. Drafting audit criteria that are binary, implementable, solution-oriented to the identified risks
4. Mapping steps #1-3 onto an ecosystem that recreates the assurance and infrastructure of trust nurtured in financial audit for more than 50 years

In support of this transformation, ForHumanity is replicating and augmenting the role of the Financial Accounting Standards Board (FASB) and the International Financial Reporting Standards (IFRS) foundation, who drafted GAAP and IFRS respectively. Unlike those predecessors, ForHumanity is a grassroots, civil-society organization with contributors from more than 98 countries around the world. Our approach ensures globally-harmonized, audit criteria that operationalize the law, standards, and best practices sourced by diverse input and multi stakeholder feedback contributors.

We draft audit criteria for AAA Systems in the context of new legislation all around the world, such as, the EU's General Data Protection Regulation (GDPR), and the EU Artificial Intelligence Act, Consumer Protection and Consumer Privacy Protection Act in the United States, Lei Geral de Proteção de Dados Pessoais (LGPD) in Brazil, and India's Digital Personal Data Protection Act

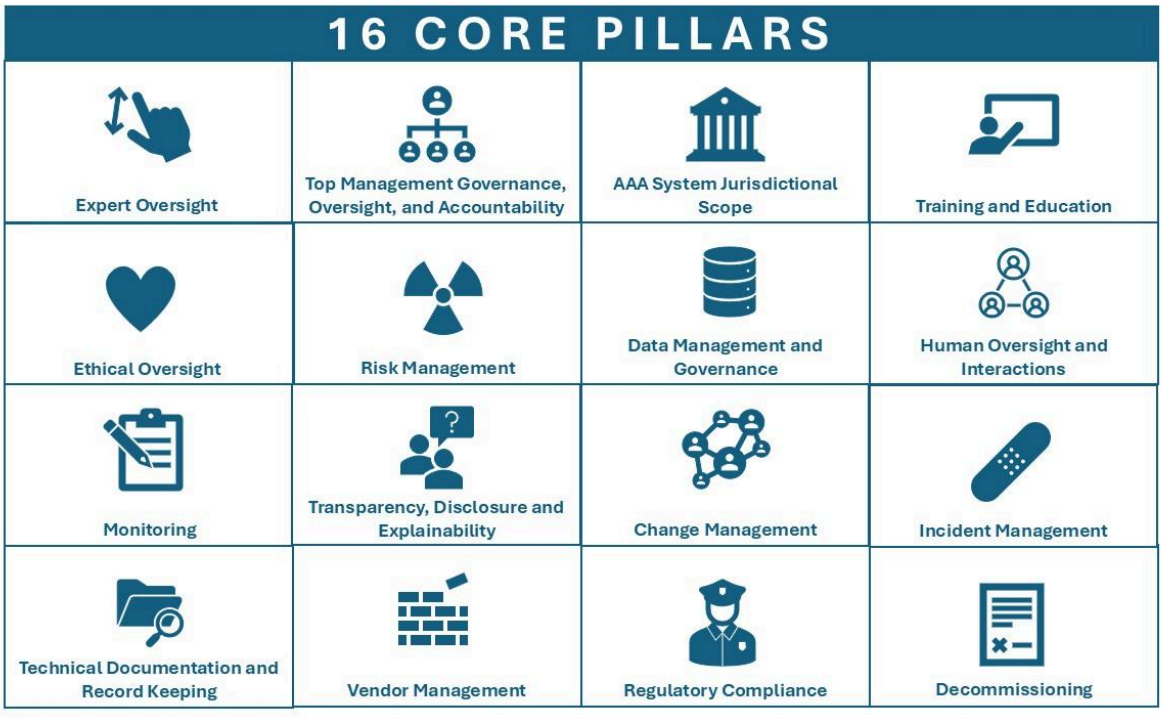
ForHumanity's authority for producing audit criteria is grounded in the robustness of our crowdsourced, transparent process (no one is excluded from participating), however we always seek the endorsement of Federal, state, and local authorities, as applicable, when they support the approval of audit criteria, such as the manner in which most nation-states and regional blocks have adopted Generally Accepted Accounting Principles (GAAP) or International Financial Reporting Standards (IFRS) to govern financial accounting and reporting. When

governments are unprepared to endorse uniform, objective audit criteria, then ForHumanity seeks adoption directly from the marketplace, which is what occurred in 1973 with GAAP and the predecessor to IFRS, prior to Federal adoption in the years afterwards.

# CORE AAA System Governance

The CORE AAA Governance certification scheme describes the foundational elements necessary for robust governance, oversight, and accountability of AAA Systems required by burgeoning legal frameworks and standards as well as the crowdsourced identification of implementable best practices necessary to mitigate risk to humans from AAA Systems.

AAA Systems are (often complex) socio-technical tools. As a result, this certification scheme is designed to ensure that globally-recognised, minimum requirements for robust AAA System governance are established and operational. This certification scheme is the cornerstone to ForHumanity's global, modular certification program. Critical compliance-by-design infrastructure is necessary for responsible and trustworthy usage of AAA Systems by corporations and the public sector regardless of jurisdiction. These elements of governance, oversight, and accountability are applicable to all AAA System operations:



# Modular Approach to Jurisdictional Compliance

To facilitate implementation of AAA System governance, oversight, and accountability. ForHumanity has established certification schemes that are modular. This CORE AAA Governance certification scheme is applicable for all AAA Systems that are not considered to be low risk and represents the non-negotiable foundational requirements necessary to engage in robust governance, oversight, and accountability of AAA Systems. The elements established as a result of compliance with this certification scheme are fully integrated with all ForHumanity modular certification schemes, such as GDPR, EU AI Act, Unfair, Deceptive, or Abusive Practices, Cybersecurity, or the Children's Code.

As a result of this modular process, for any given AAA System, multiple certification schemes may be applicable to assure the auditee of compliance with all applicable Relevant Legal Frameworks and certification schemes. For example, an AAA System that is a Provider of a recommendation system for applicant screening is an annex III high-risk, automated employment decision tool, that processes Personal Data - in such a circumstance the AAA System would likely need all of the following modules to ensure comprehensive compliance with all applicable legal obligations in the European Union:

1. ForHumanity CORE AAA System Governance Provider-only certification scheme (Foundational governance)
2. ForHumanity EU GDPR Controller-Only certification scheme (GDPR Compliance - Personal Data)
3. ForHumanity EU AI Act certification scheme (Annex III - high-risk)
4. ForHumanity Global Disability Inclusion and Accessibility certification scheme (EU 2019/882 - European Accessibility Act of 2019)

Assurance will be granted to an auditee based upon the certification scheme for which they are currently compliant with according to their auditor. An auditee at the initial phases of building certification under the ForHumanity process might find itself certified only under the CORE scheme, while they are conducting their compliance audits on additional modular schemes. The auditee is entitled to disclose all Independent Audit of AI Systems certification scheme seals (and associated disclaimers) for which it is currently compliant.

# Assessment of Modular Certification Schemes

Providers have a duty to comply with all legal obligations associated with the deployment of their AAA Systems. ForHumanity supports the compliance of these obligations by establishing certification schemes for laws, regulations, and standards applicable to AAA Systems. An

auditee may determine which modular certification schemes provide comprehensive assurance, by assessing the AAA System to determine which additional certification schemes are appropriate and applicable. This assessment includes the following considerations (but not limited to):

1. In what Jurisdictions does the AAA System operate and what are the applicable Relevant Legal Frameworks? Example ForHumanity certification schemes include:
    a. EU AI Act
    b. Digital Services Act
    c. Consumer Protection
    d. Children's online safety laws
        i. UK Children's Code
        ii. Age-Appropriate Design Code
    e. Equality and nondiscrimination law in regards to accessibility and inclusion
        i. US Disability Inclusion and Accessibility
        ii. EU Disability Inclusion and Accessibility
        iii. Canada Disability Inclusion and Accessibility
    f. UK Consumer Duty
    g. SM & CR
    h. NYC AEDT Bias Audit certification scheme
    i. State or Local certification schemes offered by ForHumanity
2. Is Personal Data present in the AAA System? Example ForHumanity certification schemes include:
    a. GDPR (EU and UK)
        i. Controller,
        ii. Processor (Integrated)
        iii. Processor (Standalone)
    b. DPDPA (India)
    c. CCPA (California)
    d. LGPD (Brazil, English and Portuguese)
    e. DIFC and Reg 10 (Dubai)
    f. PIPA (Bermuda)
3. Is the AAA System covered by a ForHumanity use case specific certification scheme, example ForHumanity certification schemes include:
    a. Automated Employment Decision Tools
    b. LLM/LMM/Digital Worker
    c. Model Risk Management
4. Is the AAA System sufficiently cybersecure according to any applicable legal requirements (e.g., NIST Cybersecurity 2.0, ForHumanity Cyber Security, Regulation (EU) 2024/482)?

5. Would the auditee want to certify portions of its AI Governance, Oversight, and Accountability? Example ForHumanity certification schemes include:
    a. Top Management and Oversight Bodies
    b. Ethics Committee
    c. Risk Management

Providers will increase trust and marketability with Deployers, limit their liability, and control their risk of legal compliance by achieving assurance with all Relevant Legal Frameworks.

# 1.0 Scope

ForHumanity designed this certification scheme for Providers (Auditees) of any size. The scheme may be applied to one or more specific AI[2], Algorithmic, or Autonomous Systems (including General-Purpose AI), however it may not be used for AAA Systems that are prohibited under the AAA System's operating jurisdiction. The certification scheme is valid for 12 months unless significant changes occur (see Audit Period of Validity - in the ForHumanity Audit Manual).

Establishing "if" the AAA System is in scope for this certification scheme requires a scope assessment that goes through the following steps to determine applicability:

1) Using ForHumanity's Low Risk Assessment, determine whether the certification scheme is being applied on a voluntary or mandatory basis
2) Assess whether the Target of Evaluation (as defined in Section 1.2) falls under the definition of AI, Algorithmic, or Autonomous Systems (definitions found in section 3.0)
3) In regards to the jurisdiction of deployment, assess whether the AAA System is prohibited in the jurisdiction of assurance for any applicable certification scheme

## 1.0.1 Determination of Provider or Deployer Status

Prior to engaging with this certification scheme, it is necessary for the organization that will be the auditee to verify that they are a Provider of the AAA System in order to use this scheme effectively.

An organization is defined as a Provider of an AAA System if any of the following are true:

*natural or legal person, public authority, agency or other body that:*

[2] Artificial Intelligence - Autonomous machine or software, that may learn and that replaces a function or task related to human decision-making

1. *Develops an AAA System or*
2. *Sells, shares, or trades the AAA System for deployment or*
3. *Any customization (including fine-tuning and optimization) of an AAA System, including placing its brand or trademark on the AAA System*

An organization is defined as a Deployer of an AAA System if all of the following are true:

*natural or legal person, public authority, agency or other body that:*

1. *Acquires, configures, and operates an AAA System within existing parameters, specifications, terms and conditions as defined by a Provider and*
2. *Does not operate an AAA system with its own brand or trademark*

This certification scheme requires Providers (Auditees) to identify all applicable jurisdictions in which the AAA System operates in order to determine additional applicable legal obligations. These additional legal obligations are called Relevant Legal Frameworks (a defined term). Relevant Legal Frameworks are assessed and documented in Section 5.0 criteria. Deployers should use the CORE AAA System Governance Deployer-only v1.5 scheme.

## 1.1 Out of Scope Systems

Systems that are not AI, Algorithmic, or Autonomous Systems, according to the definitions found in Section 3.0, are out of scope.

AAA Systems placed on the market by Deployers are out of Scope for this certification scheme, however ForHumanity offers a separate certification scheme for Deployers.

AAA Systems that are currently prohibited, based upon a crowdsourced interpretation of legal prohibition from a collection of countries from around the world including the following list are out of scope:

1. Emotion Recognition Systems (according to the EU AI Act)
2. Social Scoring (according to the EU AI Act)

ForHumanity will examine legally prohibited systems in any of the following Jurisdictions (United States, Canada, Brazil, Australia, New Zealand, Singapore, Japan, India, United Arab Emirates, South Africa, European Union, and the United Kingdom) periodically, to reassess and add or remove AAA Systems from the out of scope list.

## 1.2 Target of Evaluation Determination Process

The organization seeking certification determines the AAA System to which the certifying body will apply the scheme and documents this agreement in a contract. The Target of Evaluation (ToE) shall be defined by contract between the certifying body and the organization. The certification is valid for 12 months from the date certification is issued by the certifying body.

The contract shall document all information required by the certifying body for a sufficient Certification Plan and shall include all of the following:

1) Name/identifier of the ToE, specifically noting all inputs and outputs of the **AAA System** as described in the **System Architecture Report** - Document that describes the overall, top-level blueprint of conceptual/logical/physical structure of the system including relevant frameworks and applicable standards (e.g., ISO, CEN/CENELEC, IEEE) and includes descriptions of **Processor**, and sub-**Processor** relationships including databases, processing, flow and movements, pipeline, data collection, UX interfaces, and location/**Jurisdiction** and the **Data Flow Diagram.**
2) Systems or organizations expected to be "in" or "out" of scope (including a visual representation as appropriate). "In" and "out" of scope applies to third parties (including Processors) under contract.
3) The AAA System will be specifically identified including its Scope, Nature, Context, Purpose. For "out" of scope adjacent or interdependent processing or systems shown in the **System Architecture Report**, the organization shall document and justify "out" of scope boundaries for those adjacent or interdependent processing or systems.
4) Description of the data deployed in the system, specifically noting the Personal Data and Sensitive Data that may be present (including Inferences and/or potential Proxy Variables).
5) Specify where the processing of data happens in terms of physical location, including whether or not there are transfers to third countries or international organizations and whether such transfers are a part of the ToE or out of scope to the ToE.
6) Identify a process for assessing all applicable jurisdictions in which the AAA System operates in order to determine additional applicable legal obligations, called Relevant Legal Frameworks.

The certifying body will only perform an audit of the documented scope. The Provider bears the responsibility of ensuring that all necessary components of the AAA System are covered in the definition of the ToE.

The ToE shall be defined in such a way that it is not misleading or likely to be misinterpreted by third parties.

The ToE may include elements of the application that are NOT AAA Systems themselves but are necessary to ensure that the AAA System functions according to the defined Scope, Nature, Context, and Purpose. This certification scheme is NOT limited to certifying only the AI, Algorithmic or Autonomous component, but rather the entirety of the AAA System application, sometimes referred to as "AAA System and supporting ecosystem".

## 1.3 Territorial Scope

This is a global certification scheme and applies to AAA Systems that impact AI Subjects regardless of the jurisdiction.

# 2.0 Normative References

ISO 27001/27002:2022 - Information security management
ISO 15288:2015 - Systems and software engineering — System life cycle processes
ISO 31000 - Risk management
ISO 9001 - Quality Management Systems
ISO 9001:2015 - Quality Management Systems
ISO/IEC 25010
ISO/IEC 25059

# 3.0 Terms and Definitions

Defined terms are bolded and capitalized throughout this document.

| Defined Term | Definition |
|---|---|
| **AAA Cybersecurity Lead** | An expert accountable for security and cybersecurity policies, processes, procedures, risk management, incident response, business continuity and disaster recovery |
| **AAA System** | Any end-to-end application containing an AI, Algorithmic, or Autonomous component including both technical elements (e.g., databases, data, networks, hardware) and lifecycle elements (e.g., pre-processing, monitoring, human oversight) |
| **AAA Systems List** | A list, either by name or other identifier that tracks all distinct AI, algorithmic or Autonomous Systems |

| | |
|---|---|
| **AAA System Deployer Guide** | A digital documentation that intends to enable and empower the Deployer with information about the **AAA Systems** from the **Provider** that is necessary to successfully operate the **AAA System**. |
| **AAA System AI Subjects Guide** | A digital documentation that intends to enable and empower the **AI Subject** with information about the **AAA Systems** from the **Provider** or **Deployer** that is necessary to successfully operate the **AAA System**. It is digital information that is concise, complete, correct, clear, relevant, accessible and comprehensible to the **AI Subject** |
| **Accessibility** | degree to which a product or system can be used by people with the widest range of characteristics and capabilities to achieve a specified goal in a specified context of use [SOURCE: ISO/IEC 25010:2011] |
| **Accessibility Conformance Report** | Evaluation, overseen by the **Disability Inclusion and Accessibility Committee**, that determines how well **AAA System** and associated ecosystem meet accessibility standards (e.g., VPAT EN 301 549, Model Accessibility Statement, GPAT, Open ACR) https://www.w3.org/WAI/test-evaluate/conformance/ |
| **Accommodation** | a timely adjustment made in a system (such as the provision of tools or changes to the environment or the way in which the **AAA System** is usually provided) to accommodate or make fair the same system for individuals, including **Persons with Disabilities** based on a need, which will likely vary. **Accommodations** can be religious, physical, mental or emotional, academic, or employment related and are often mandated by law and jurisdictionally sensitivity |
| **Accuracy** | The closeness of agreement between a test result and the accepted reference value SOURCE: ISO 3534-1 [A measure of a system's **Functional Correctness]** |
| **Adaptability** | degree to which a product or system can effectively and efficiently be adapted for different or evolving hardware, software or other operational or usage environments [SOURCE: ISO/IEC 25010:2011] |

| | |
|---|---|
| **Adverse Impact** | When the selection rate of a Protected Category is below 4/5th or 80% of the highest selection rate |
| **Adverse Incidents** | negative outcomes or impacts to natural person caused by AAA System |
| **Adverse Incident Reporting System (AIRS)** | A system available to the Consumer/Customers/Users/AI Subjects (including internal stakeholders, partners, customers, civil society, industrial associations, and the general public) to report or register confidentially (and maturely anonymously) information regarding their perceived or realised adverse incidents attributable to AAA Systems |
| **Age-Appropriate** | A commitment to delivering suitable information, content, services, applications, interfaces and design that considers a Child's specific developmental stage (capacity, skills and behaviours) and understanding, leading to beneficial engagement that supports the well-being of the Child. Age-Appropriate content and disclosure includes the identification of target age ranges. This includes signposting for when a child is instructed to seek parental assistance. Notifications shall be in plain language |
| **AI Compliance Lead** | A natural person assigned to be responsible for leading regulatory compliance functions including acting as a Point of Contact for communications with certification bodies and supervisory authorities |
| **AI Subject** | A natural person who is impacted by the outcomes of a AAA System |
| **Algorithm** | a process or set of rules to be followed in calculations or other problem-solving operations, especially by a computer |
| **Algorithm Ethics** | A sub field of Ethics focused on instances of Ethical Choice emerging from AI, algorithmic and autonomous systems. Training and expertise include areas such as Necessity, Proportionality, Benchmark setting, Validity, reliability, Model, Data and Concept Drift and thresholds for Bias mitigation. |

| | |
|---|---|
| **Algorithmic Risk** | Any risk input or indicator identified in the Algorithmic Risk Assessment, exclusive of security and cybersecurity risks inputs and indicators |
| **Algorithmic Risk Assessment** | An analysis of all risks associated with the comprehensive lifecycle of an AAA System, not covered by the Cybersecurity Risk Assessment, the Ethical Risk Assessment, the Committee Governance Assessment and the Systemic Societal Impact Analysis. |
| **Algorithmic Risk Committee** | group of employees (or outsourced expert group) tasked with assuring that all AI, algorithms and autonomous systems have taken the necessary steps to identify, remediate, mitigate, explain, monitor and document all instances of Algorithmic Risk |
| **Architectural Inputs** | parameters, variables, hyperparameters, weights and other elements that are used to establish an algorithmic calculation or process |
| **Artificial Intelligence** | a process or system that replaces human decision-making |
| **Authenticity** | degree to which the identity of a subject or resource can be proved to be the one claimed<br><br>[SOURCE: ISO/IEC 25010:2011] |
| **Authorised Representative** | a natural or legal person located or established in a Jurisdiction who has received and accepted a written mandate from a provider of an AI system or a general-purpose AI model to, respectively, perform and carry out on its behalf the obligations and procedures established by applicable Relevant Legal Frameworks |
| **Authority** | The legal right to hold or provide data |
| **Automation Bias Curriculum** | A body of learning designed to raise awareness of the Human-in-Command and other employees associated with the AAA System in regards to a general over-reliance of AAA Systems. The curriculum is designed to establish a healthy scepticism in regards to AAA Systems and to educate users when AAA Systems can be relied upon and when they should |

| | |
|---|---|
| | be stopping, pausing, disregarding, overriding, and reversing. The curriculum further encourages users to acquire knowledge and understanding of underlying assumptions, data inputs, risk mitigations, and Residual Risk associated with the AAA System. |
| **Autonomous System** | Any self-governing system, operating without a human-in-the-loop (excluding pre-start inputs and design plus maintenance, recalibration, retasking and repair) , producing characteristics of human dexterity, such as arm or leg motion and their results (e.g., travelling distances) or any one of the five human senses |
| **Availability** | degree to which a system, product or component (including data) is operational and accessible when required for use<br><br>[SOURCE: ISO/IEC 25010:2011] |
| **Bias Risk Assessment** | Subset of an **Algorithmic Risk Assessment**, still including **Diverse Inputs and Multi Stakeholder Feedback** (DI & MSF), focused on uncovering and mitigating risk associated with Bias (e.g. **Cognitive Bias** and **Technology Barrier**) in the data, architectural inputs and outcomes |
| **Biometric Data** | Personal Data resulting from specific technical processing relating to the physical, physiological, or behavioural characteristics of a natural person, including DNA, (e.g., imagery of the iris, retina, fingerprint, face, hand, palm, vein patterns, and voice recordings, keystroke patterns or rhythms, gait patterns or rhythms, sleep data, health data, or exercise data) that can be used to uniquely identify that natural person<br><br>*a subset of sensitive or Special Category data |
| **Bug Bounty Program** | A way for a Provider to reward an external security tester for identifying flaws, vulnerabilities, errors and bias in an AAA System and reporting them to the Provider through a predetermined mechanism and process |
| **Business Rationale Report** | In the context of the **Fundamental Rights Impact Assessment**, **Proportionality Study** and **Necessity** |

| | |
|---|---|
| | **Assessment**, document the system's underlying logic, **Causal Hypothesis**, **Construct Validity**, and feature relevance that upholds and supports the human rights and freedoms |
| **Business Continuity Plan (BCP)** | scheme that describes a system of prevention and recovery from potential threats to a company, ensuring that personnel and assets are protected and are able to function quickly in the event of a discontinuity, threat or disaster. The BCP is integrated with a Contingency plan and restoration prioritisation plan |
| **cAIRE report** | Comprehensive Artificial Intelligence Risk Evaluation report, comprising all risk inputs, risk mitigations and Residual Risks gathered from any of the following reports: Algorithmic Risk Assessment, Systemic Societal Impact analysis, Test Completion Report, Ethical Risk Assessment, and Committee Governance Assessment |
| **Capacity** | degree to which the maximum limits of a product or system parameter meet requirements [SOURCE: ISO/IEC 25010:2011] |
| **Causal Hypothesis** | An assessable proposition, to be proven or disproven, that predicts a relationship between two variables, where the change in the first variable brings about change in the second variable |
| **Change Management Impact Assessment** | An assessment of metrics, measurements, and thresholds pre-determined to delineate between minor changes that can be classified as version updates versus major changes that represent meaningful risk to the organisation, Deployers, or AI Subjects |
| **Change Management Plan** | An ISO 9001:2015 document that delineates implementation, risk, impact, and adaptation/migration strategies as well as communication procedures, procedures for unplanned outages from change, processes for protection of production data, defined backup and rollback procedures, and supporting documentation for approval |
| **Child (ren)** | a person under the age of 18 |

| | |
|---|---|
| **Child's Data Oversight Committee (CDOC)** | A group of 3 or more people which may comprise outside experts, tasked with reviewing all aspects of data collection, risk and procedures associated with data related to Children for the Jurisdiction |
| **Child-Friendly** | To present information using diagrams, cartoons, graphics, video and audio content, and gamified or interactive content that will attract and interest Children, rather than relying solely on written communications |
| **Choice Architecture** | The inputs to a recommender system that may be controlled or modified by the AI Subject |
| **Code of Data Ethics** | set of guidelines, principles and procedures by which data is acquired, analysed, processed, adjusted, compiled or otherwise sold, traded or shared with other entities |
| **Code of Ethics** | a Publicly documented set of principles and rules concerning moral obligations and regards for the rights of humans and nature, which may be specified by a given profession or group. The document is drafted and kept up to date by an organisation's Ethics Committee and outlines said organisation's shared moral framework within the Relevant Legal Frameworks, providing context to instances of Ethical Choice, diversity and anti-discrimination |
| **Cognitive Bias** | The way a particular person understands events, facts, and other people, which is based on their own particular set of beliefs and experiences and may not be accurate in regards to the Data Subject or sample population resulting in discriminatory outcomes (e.g., confirmation bias, anchoring bias) |
| **Committee Governance Assessment** | An analysis and designation of accountability, oversight and responsibility for committees (Ethics Committee, Algorithmic Risk Committee, and specialty committees such as the Children's Data Oversight Committee, Disability Inclusion and Accessibility Committee), designated individuals (per a Duty Designation Letter), the Chief Executive Officer and the Board of Directors for any/all risk associated with an AI, algorithmic |

| | |
|---|---|
| | or autonomous system including duties associated with compliance with audit criteria |
| **Concept Drift** | The change in the measured relationships (e.g., correlation, covariance) between input and output data resulting in misalignment |
| **Confidentiality** | degree to which a product or system ensures that data are accessible only to those authorised to have access<br><br>[SOURCE: ISO/IEC 25010:2011] |
| **Consent** | means any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of Personal Data relating to him or her |
| **Construct validity** | How well a set of indicators represent or reflect a concept that is not directly measurable. The extent to which feature (indicator) relevance, **Functional Correctness**, and causality of a model or algorithm represent the ground truth with the theoretical construct |
| **Context** | The circumstances in which an event occurs; including jurisdiction and/or location, behaviour and functional inputs to an AAA System that are appropriate (e.g. domain, operating environment) |
| **Contingency Plan** | A plan to make the system inaccessible and unavailable, or to continue processing, in the context of a security related event. |
| **Controllability** | degree to which a Provider, Deployer and/or AI Subject can appropriately intervene in an **AAA System's** functioning in a timely manner<br><br>Modified from the ISO definition [SOURCE: ISO/IEC 25059:2023] |
| **Corrective Action Plan** | Summarizing responses to a Accessibility, Change Management Plan, Incident Response, including containment, eradication, recovery, and implementation of new risk controls, treatments, and/or mitigations |

| | |
|---|---|
| **Cybersecurity Risk Log** | A separate and secure risk log that contains risk inputs and indicators in regards to security and cybersecurity risks and vulnerabilities |
| **Data Age** | The elapsed time between the original acquisition or compilation of each datum and current state |
| **Data Curation Report** | Documenting the evaluation of the system's sample data and describing the data collection process including **Availability**, quantity, suitability, **Provenance**, sampling method, Ground Truth **Availability**/verifiability. Documenting the sample data preparation process, including the treatment of anomalies and exceptions, specifying data cleaning, encoding, transforming, enriching and aggregating tasks. |
| **Data Drift** | occurs when the distribution of incoming data to a model changes over time, or differs from the data used to train and test the model resulting in misalignment |
| **Data Evaluation Report** | Documenting a technical understanding of the sample data, including describing, in detail, the statistical characteristics (syntactic metadata) of the sample data (e.g., range, mean, median, mode, missing values, volatility, shape, modality, ratio of features values, data format). Documenting that the sample data is sufficiently diverse, balanced, and representative including all bias mitigation s |
| **Data Entry Point Attacks** | vulnerabilities and attacks associated with the data used for training and processing data, where the adversary manipulates the data in order to attack, alter or otherwise corrupt the intended purpose, scope and nature of the algorithmic system (e.g., **Data Poisoning**, model inversion, model evasion) |
| **Data Flow Diagram** | Picture or graphic which visually represents all inputs and outputs of data (e.g., Personal and Non-Personal Data) associated with an AI, Algorithmic or Autonomous (AAA) System across controller and processor relationships including databases, processing, flow and movements, pipeline, data collection, UX interfaces, location/Jurisdiction |
| **Data Integrity** | A property possessed by data items that have not been altered |

| | |
|---|---|
| | in an unauthorised manner since they were created, transmitted, or stored (source: NIST) |
| **Data Lead** | An expert accountable for Data Management and Governance of an **AAA System** |
| **Data Poisoning** | An adversarial attack targeted at training, testing/validation, Data Quality, Information Quality, Pipeline Data in an attempt to render the data useless or alter/damage the model's ability to achieve its scope, nature, context and purpose potentially altering outputs in favour of the adversary. Intentional subversion of Data Quality |
| **Data Protection Impact Assessment** | to assess the data protection risks related to the processing of Personal Data |
| **Data Quality** | Data that is expected to be fit for purpose, representative, and aligned to the Scope, Nature, Context and Purpose of the intended use as applicable to an AAA System. Data Quality is characterised as complete, accurate, categorically representative, consistent, precise collected from reasonably calibrated sensors, surveys, or other tools to gather data |
| **Data Subject** | means an identifiable natural person who can be identified, directly or indirectly, in particular by referencing an identifier such as a name, an identification number, location data, an online identifier, or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural, or social identity of that natural person. |
| **Data Transparency Document** | A clear and plain language, public report created by the **Algorithmic Risk Committee** designed to collect and document all relevant steps taken by the **Ethics Committee** and the **Algorithmic Risk Committee** to mitigate risk of **Bias**, insufficient **Data, Information**, and **Pipeline Quality** |
| **Dataset** | collection of datum intended for use in AI, algorithmic or autonomous systems<br>1) Raw - capture of data, prior to any manipulation (e.g. cleaning, labelling, organising) as acquired from any source in its original form |

| | |
|---|---|
| | 2) Structured - collection of data, post enhancement by cleaning, labelling, organising) |
| **De-Anonymization** | the practice of willfully processing aggregate anonymized data for the purposes of re-identifying individual persons specifically regardless of the rationale or merit. De-Anonymization does not include security tests which are designed with the knowledge of the owner and anonymizer of the aggregate data to verify the quality and security of the anonymization process |
| **Deceptive Design** | encompasses "dark patterns", coercion, conditioning and subliminal behaviour modifications |
| **Decommissioning Policy** | A document that identifies all of the process, procedures, metrics, measurements and thresholds that might lead to decommissioning of the AAA System, including the process for decommissioning |
| **Deletion** | (delete) in the context of data, is when data is removed and is no longer available in plain sight or can easily be recovered |
| **Deployer** | a natural or legal person, public authority, agency or other body that acquires, configures, and operates an AAA System within existing parameters, specifications, terms and conditions as defined by a Provider and does not operate an AAA system with its own brand or trademark |
| **Deployer and AI Subject Contract log** | A log of contractual relationships between the **Provider** and a **Deployer** or **AI Subject** including required deliverables per the contract (related/integrated to **AAA Systems List** and **Identity and Access Management log**) |
| **Destruction** | (destroy) in the context of data, is when data is removed from your device and can never be restored, even be professional data recovery experts |
| **Disabled Person(s)/ People with Disabilities** | include those who have long-term physical, mental, intellectual or sensory impairments which in interaction with various barriers may hinder their full and effective |

| | |
|---|---|
| | participation in society on an equal basis with others |
| **Disability Inclusion and Accessibility Committee (DIAC)** | A group of 3 or more people which may comprise outside experts, tasked with reviewing all aspects of data collection, risk and procedures associated with data related to Persons with Disabilities for the Jurisdiction |
| **Diverse Inputs and Multi Stakeholder Feedback** | As accepted by the **Ethics Committee** in compliance with the **Code of Ethics** and/or a diversity policy, it is a collection of individuals noteworthy by their representation of lived experiences, backgrounds, cultures, diversity of thought processes, skills, expertise (including domain experts), and inclusion of **Protected Categories** and **Intersectionalities**. This group is used for risk inputs, risk evaluation, assessment of foreseen misuse and this evaluation occurs throughout the algorithmic lifecycle from design to decommissioning (captured in an **Algorithmic Risk Assessment**) |
| **Emergent Risk** | An unforeseen risk that was not contemplated and is presently manifest. A risk where the potential for harm or loss is not fully known presently. |
| **Ethical Choice** | For a natural person, an ethical choice is the result, outcome or judgement made using a shared moral framework, or set of moral principles based upon the organisation's Code of Ethics. It requires awareness and consideration of a set of options to be made in the context of Artificial intelligence, algorithmic or autonomous systems, using a set of principles and rules concerning moral obligations and regards for the rights of humans and for nature, which may be specified by a given profession or group |
| **Ethical Choice Curriculum** | Body of learning designed to raise awareness of instances of Ethical Choice for designers, developers, governance and oversight teams involved in the creation of AI, algorithmic and autonomous systems. The curriculum raises awareness of instances of Ethical Choice as well as the organisation's preferred procedure for handling the instance of Ethical Choice. |
| **Ethics Committee** | A group of persons trained in Algorithm Ethics and Ethical |

| | |
|---|---|
| | Choice, guided by the Code of Ethics and Code of Data Ethics, which they create and maintain on behalf of the organisation. The Ethics Committee is responsible for all instance of Ethical Choice related to AI, algorithmic and autonomous systems and producing the Ethical Risk Assessment |
| **Ethical Risk Assessment** | A study of instances of Ethical Choice, softlaw, application of Code of Ethics and Code of Data Ethics principles and shared moral frameworks across the lifecycle of the AI, algorithm or autonomous systems shared Publicly. |
| **Event Log** | An event is any activity carried out by the system(e.g., request for data, remote login, automatic shutdown of the system, or deletion of a file) or through an interaction with the system.<br><br>The Event Log must contain five main components:<br><br>• user ID<br>• System activity can be monitored to identify what took place.<br>• At a certain date and time, an event occurred.<br>• The event took place on the device/system and its location was identified.<br>• Network addresses and protocols – IP information.<br><br>Paraphrased from [ISO 27001:2002] |
| **Exceptions Interpretability** | timely interface designed for human oversight during the period in which the **AAA System** is in use for identification of:<br>A. Anomalies,<br>B. Dysfunctions,<br>C. Exceptions,<br>D. Expected foreseeable misuse,<br>E. False positive and false negative<br>F. **Key Risk Indicators** (KRIs)<br>to enable and empower a **Human-in-Command to** stop, pause, disregard, override, and reverse the **AAA System** |

| | |
|---|---|
| **Explainability Statement** | A description of the AAA System, its logic and any applicable automated decision-making, including profiling (inferences), when the outcome impacts the health, safety, and human rights of an AI Subject that sufficiently describes the model in plain language in order to provide understanding to the AI Subject on how conclusions were reached both globally and in the context of a specific case (locally) |
| **Explainability+** | A human-centric process by which an AI Subject is helped to understand the decision making process and educated on how they could have earned a favourable result from the system, in order to improve their interaction, their outcome or their satisfaction |
| **Failure Mode and Effects Analysis** | A methodology for collecting knowledge about possible points of failure in a design, process, product, or service |
| **Functional Correctness** | degree to which a product or system provides the correct results with the needed degree of precision [SOURCE ISO/IEC 25010:2011, 4.2.1.2] |
| **Fundamental Right Impact Assessment** | An analysis of the manner in which an **AAA Systems** interacts with the rights and freedoms guaranteed to **AI Subjects** according to the **Relevant Legal Frameworks** |
| **Geolocation** | process of finding, determining and providing the exact location of a computer, phone, tablet, networking device or equipment, and may including inputs such as wifi, IP Address, bluetooth connectivity, GPS, latitude, longitude, altitude, direction of movement and time period recorded |
| **Ground Truth** | Information ascertainable as real or true through observation or experience |
| **Guardian** | person of legal age and ability who can act on behalf of an Minor, Child, or Person with Disability |
| **Human Interactions Report** | this report tracks all human interactions, their effectiveness and impact on a **AAA System** |

| | |
|---|---|
| **Human Interactions Log** | This records all **Human-in/on-the-Loop/Command** interactions with the high risk **AAA System**, including measures implemented during interactions |
| **Human-in-Command** | A natural person assigned by a system Provider or Deployer to act as no less than a 3rd line of defence governance Human-on-the-Loop, knowing the **Capacity** and limitations of the system, possessing sufficient training for the regular operation including the identification of anomalies, dysfunctions and unexpected performance. |
| **Human-in-the-loop** | any model that is unable to offer an answer or conclude processing without human intervention |
| **Human-on-the-loop** | Human supervision and/or control of AI, algorithmic or autonomous systems, however the system is able to conclude processing without the need for human intervention |
| **Inclusivity Risk Assessment** | A process examining Training Data, designed to identify risk inputs associated with bias, inclusivity, accessibility, safety and security of AAA Systems, and further identify treatments and mitigation. The testing examines the potential for adverse incidents associated with AAA Systems when tested with extreme examples (including black swan, fat-tail, boundary values, failure of expected inter-relationships etc) and "Edge-in" thinking designed to balance the innate nature of most algorithms to "normalise" or find the "best fit" |
| **Inference** | assumption or conclusion reached by a data processing algorithm, which may not be treated as fact and shall be labelled as such. |
| **Information Quality** | Data Quality that has demonstrated fitness for purpose, representative and aligned to the **Scope, Nature, Context** and **Purpose** of the intended use as applicable to an **AAA System**. Information Quality is characterised by **Construct Validity**, **Provenance**, **Authority, Authenticity, Relevance,** and **Data age**, legal basis and **Consent,** if applicable |
| **Integrity** | degree to which a system, product or component prevents |

| | |
|---|---|
| | unauthorised access to, or modification of, **AAA Systems** and/or data<br><br>Modified from [SOURCE: ISO/IEC 25010:2011] |
| **Internal Independence** | A requirement that natural persons assigned to assess or validate any of the following:<br>1. specifications,<br>2. requirements,<br>3. processes,<br>4. procedures, or<br>5. Implementations<br>were not involved in the design, development, data curation, or implementation of the assessed or validated activities |
| **Internal Sign Off Report** | Assess whether the system is fit to be deployed, legally, ethically and consistently with the original **Causal Hypothesis**, especially in consideration of impacts to health, safety, and fundamental rights. |
| **Intersectionalities** | The places, ways, and sources of inequality in systems based on combinations of gender, race, ethnicity, sexual orientation, gender identity, disability, class, and other forms of discrimination to create unique dynamics and effects. A subset of categories of **Protected Categories.** |
| **Intervenability** | degree to which an operator can intervene in an AI system's functioning in a timely manner to prevent harm or hazard<br>[SOURCE: ISO/IEC 25059:2023] |
| **Jurisdiction** | a defined geographic area over which a particular legal authority may lawfully exercise control |
| **Just-in-Time** | The moment a notification is presented to the AI Subject prior to an interaction with the AAA System that could be any of the following:<br><br>A. A statement of rights (e.g., Disability Inclusion and Accessibility Statement) |

| | |
|---|---|
| | B. A legal obligation (e.g., the collection of Personal Data)<br>C. terms and conditions |
| **Key Detrimental Indicators** | Parameterised content, where the content, regardless of medium (e.g., AR/VR, audio, images, video, profile/comments, etc), is determined to be any of the following:<br>**A. Illegal Content (e.g., Terrorism, Child Sex Abuse Material, Hate Speech,** discriminatory**)**<br>B. harmful or negatively impacting to the well-being of recipient of the service, including **Vulnerable Populations,** such as:<br>a. **Adult Content**<br>**b. Bullying**<br>c. Defamatory/Slander/Libellous content<br>d. Misrepresentation and identity fraud<br>e. indications of self-harm, suicide, violence, and disorders<br>f. intentional censorship designed to circumvent monitoring (e.g., F***, S*!T)<br>g. Representations of any of the aforementioned items (e.g., emojis, GIFs)<br>h. Via goods or services (e.g., Spam, Malware, illegal goods, fraud)<br>C. **Disinformation**<br>D. restricted by guidelines and codes of practice<br>E. Breach of copyright and other intellectual property rights<br>where the parameters are subsequently deployed for monitoring and measuring in order to censor, filter or restrict the content in the **ISS** |
| **Key Language Indicators** | Parameterised content, in the context of **ESCO, ONET,** and Protected Categoriness (e.g., Ageism, Racism, Genderised, Ableism terms), that are designed to identify and remediate word choices, from an LLM, LMM, or questionnaire that could otherwise bias the reader and lead to a discriminatory outcome |

| | |
|---|---|
| **Key Performance Indicators (KPIs)** | measurements indicated in advance to determine the success or failure of an algorithmic model to achieve its purposes |
| **Key Regulated Product Indicators** | Parameterised content, where the content regardless of medium(e.g., AR/VR, audio, images, video, profile/comments, etc), is related to a product or service and is determined to be any of the following:<br>A. **Illegal Content (e.g., Terrorism, Child Sex Abuse Material, Trafficking)**<br>B. Fraudulent, or infringing of a copyright or intellectual property<br>C. Regulated Goods, sold or displayed improperly (e.g., weapons, alcohol, tobacco, pharmaceuticals, illicit drugs)<br>D. Regulated Services (e.g., Financial products)<br>E. Adult Content and age-restricted sales of goods and services<br>F. Regulated Goods biologics (e.g., bacteria, virus, fungus, livestock, plants and seeds, birds, fish and sea creatures)<br>G. Illegal cultural appropriation<br>H. Regulated Military or Defence industry items<br>I. Regulated or illicit services<br>J. Monetary items and assets<br>K. Regulated electronics, code, or technology (e.g., software or hardware)<br>L. Online gambling |
| **Key Risk Indicators (KRIs)** | Measurements and thresholds of model health and fitness that identify any of the following:<br>A. A realised deviation based upon any of the following:<br>i. **Pipeline data** that fails to conform to the data scheme<br>ii. **Exceptions Interpretability** outputs identifying anomalies, outliers, or exceptions<br>iii. **AAA System** output that exceeds thresholds<br>iv. **Adverse Incident Reports** that exceed thresholds,<br>B. A new, **Emergent Risk** |

| | |
|---|---|
| **Low Risk Assessment** | An evaluation, following the ForHumanity Risk Management Framework, to determine if the AAA System needs to undergo a conformity assessment, because it is unlikely to meaningfully, directly or indirectly, negatively impact people, people groups or the human and natural ecosystem (see ForHumanity Risk Management Framework) |
| **Metadata** | information about a datum (e.g. location, owner, date, time) |
| **Model Drift** | any change (degradation or improvement) in the predictive performance of a model that results in a change to the scope, nature, context, and purpose of the model resulting in misalignment |
| **Monitoring Lead** | An expert accountable for continuous and post market monitoring of the AAA System |
| **Nature** | The forces and processes that influence and control the variables and features (e.g., foreseeable conditions, input variables) |
| **Necessity Assessment** | Produced by the **Algorithmic Risk Committee** in consultation with the **Ethics Committee,** who are guided by the **Code of Ethics** and principles portion of the **Code of Data Ethics,** to determine whether an **AAA System** is the only or best solution, considering a comprehensive set of stakeholders, in the context of the legal basis. Additionally, it analyses and determines whether the inclusion of each **Personal Datum** collected and processed by **AAA System** is vital. |
| **Novel** | Having the characteristics of being one of the following: 1) unique, 2) unprecedented, or 3) innovative and therefore possessing insufficient comparables or industry standards resulting in an unknowable risk profile |
| **Nudge (Nudging)** | Design, interfaces and notifications of an **AAA System** that leverage concepts found in behavioural economics, political theory, and behavioural sciences, to reinforce, suggest, or |

| | |
|---|---|
| | influence (consciously or subconsciously) the behaviour, actions, or decision-making of individuals or groups |
| **Penetration Testing** | Testing technique aiming to exploit security vulnerabilities (known or unknown) to gain unauthorised access |
| **Personal Data** | Any information relating to an identified or identifiable natural person ('Data Subject'); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to their physical, physiological, mental, economic, cultural or social identity. Personal Data may be a collective term encompassing specialised terms such as Inferences, Proxy Variables, Personally Identifiable Information, Personal Information, Sensitive Data, and Special Category Data |
| **Pipeline Data** | Inputs to an operational AAA System from external sources (including natural persons) via a predetermined collection mechanism |
| **Pipeline Quality** | The nature of the live data input into an operating (live) AAA System, including the manner in which the data matches to the data schema |
| **Profiling Decline** | A Deployer or AI Subject interface that allows the Deployer or AI Subject to opt out of recommendation engines or other content moderation through the use of Profiling |
| **Profile Reset** | A Deployer or AI Subject interface that allows the Deployer or AI Subject to zero-out or completely reset the Profile created by the Provider of the system for the Deployers or AI Subject interface with the AAA System |
| **Profile Re-engage** | A Deployer or AI Subject interface that allows the Deployer or AI Subject to reapply their Profile to the AAA System after a period of Profiling Decline or Profiling Reset |
| **Profiling** | Any form of automated processing of Personal Data consisting of the use of Person Data to evaluate certain aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, |

| | |
|---|---|
| | economic situation, health, personal preferences, interests, reliability, behaviour location or movements |
| **Proportionality Study** | Conducted prior to a **DPIA**, it is a documented study conducted by the **Ethics Committee** to assess tensions and **Tradeoffs** between risks to and sacrifices of the rights and freedoms of individuals or groups, balanced against the potential benefits and gains to an individual or group in the context of the **Relevant Legal Frameworks** |
| **Protected Category(ies)** | Defined under law or regulation by Jurisdiction, may include race, age, gender, religion, ability/disability, sexual orientation, creed, colour, nation of origin, socioeconomic class etc |
| **Provenance** | The history and traceability of the supply chain especially when documented or authenticated |
| **Provider** | a natural or legal person, public authority, agency or other body that develops an AI system or a general-purpose AI model or that has an AI system or a general-purpose AI model developed and places it on the market or puts the AI system into service under its own name or trademark, whether for payment or free of charge |
| **Pseudonymisation** | Processing of personal data in such a manner that the personal data can no longer be attributed to a specific Data Subject without the use of additional information |
| **Publicly** | Refers to something that is broadly available to a wide range of people outside a particular individual, company, or select group (e.g., a public-facing website, public regulatory filing, public announcement, report, advertisement, or consumer-facing document) |
| **Purpose** | The aim or goal of a system (e.g., limitations, variants) |
| **Quality Management Lead** | An expert accountable for quality functions (e.g., control, assurance, validation) for a designated AAA System |
| **Quality Management System** | A system that captures all policies, procedures and written guidance for compliance with Chapter 2 of the Act, including policies, guidance, instructions, user guides, metrics, thresholds, technical specifications, data management and |

| | |
|---|---|
| | required documentation. It also includes an operations manual for the risk management system, regulatory compliance, record-keeping, post-market monitoring and Adverse Incident Reporting Systems, communications with National and EU authorities |
| **Reasonably Foreseeable Misuse** | The use of an AI system in a way that is not in accordance with its intended purpose, but which may result from interaction with other systems or identified by human accessors for potentially negative impacts beyond the intended Scope, Nature, Context, and Purpose |
| **Relevance** | The appropriateness and meaningfulness of each datum, feature and causal hypothesis to the **Scope, Nature, Context, and Purpose** of the **AAA System** |
| **Relevant Legal Frameworks** | The collection of applicable law such as the laws that govern an entity or organisation, that govern the rights, freedoms, and privileges of a Data Subject or AI Subject, that restrict the activities and behaviours of a Provider, or put positive obligations upon an entity |
| **Reliability** | Degree to AAA System performs specified functions under specified conditions for a specified period of time. [Source ISO 25000] |
| **Representativeness** | Describes a measurement of the **AAA System** dataset, especially for a **Protected Category, Intersectionality**, and **Vulnerable Populations,** that it has comparable statistical characteristics between the training, validation, and testing datasets related to at least two different benchmarks (within a reasonable confidence level), 1) general population and 2) a reasonable explanation of the source, sample, and pipeline (target) population, with the aim of the **AAA System** dataset being reasonably similar |
| **Residual Risk** | The documented sum of all unmitigated risk pertaining a **AAA System** |
| **Resilience** | In the context of a major disruption the ability of the system to withstand and recover. The speed and capability to return to a |

| | |
|---|---|
| | sufficient level of function in accordance with the system's intended operation. |
| **Risk Appetite** | The type, amount and threshold of risk that an organisation is prepared to accept in pursuit of its strategic objectives and business plan. |
| **Risk Tolerance** | The acceptable level of variation relative to the achievement of objectives. In setting-specific risk tolerances, management considers the relative importance of related objectives and aligns risk tolerance with risk appetite. |
| **Robustness** | Degree to which an AI system can maintain its level of functional correctness under any circumstances [SOURCE: ISO/IEC 25059:2023] |
| **Scope** | The boundaries of a system, what is covered, what is not covered (i.e, in scope, out-of-scope) |
| **Sensitive Data** | **Personal Data** that reveals:<br>1. Government-issued identification including tax ID, driver's license, or passport number<br>2. Racial or ethnic origin, religious, political, or philosophical beliefs, or union membership<br>3. Account log-in, financial account, debit card, or credit card number in combination with any required security or access code, password, or credentials allowing access to an account, including transactions<br>4. Precise geolocation<br>5. The contents of mail, email, and text messages unless the Business is the intended recipient of the communication.<br>6. Genetic data<br>7. Biometric information (e.g., fingerprint, gait, retina, face print)<br>8. Physical health and mental health data<br>9. Sex life, sexual orientation, or gender |

| | |
|---|---|
| **Service Provider** | Third-party contracted provider who is supplying critical infrastructure and services to the organisation |
| **Source Data** | Set of inputs gathered from outside the model environment |
| **Statistical Bias** | Systematic and repeatable errors in a computer system that create unfair outcomes, applied specifically to **Protected Categories**, **Intersectionalities** and **Vulnerable Populations** resulting in discriminatory outcomes |
| **Synthetic Data** | Artificial data, without **Provenance**, **Authority**, or **Authenticity**, (not include data anonymization techniques e.g., masking) that is generated from original **Source Data** and represents characteristics of the original **Source Data** |
| **System Architecture Report** | Document the overall, top-level blueprint of conceptual/logical/physical structure of the system including relevant frameworks (e.g., TOGAF, Zachman) and applicable standards (e.g., ISO, CEN/CENELEC, IEEE) |
| **System Design Report** | Documents the design of the system and its design choices and associated rationale, notably in the areas of human interactions with the system, including pros and cons, tensions and **Trade-offs** amongst choices, especially in the context of **Protected Categories** and **Vulnerable Populations** and log design choices of the overall system at various levels of granularity based upon the complexity of the system |
| **System Development Report** | Document the development process followed in building the system including approach, phases, methods (CRISP-DM, SDLC), techniques, procedures, and tools |

| | |
|---|---|
| **Systemic Societal Impact Analysis** | A study designed to consider, track and measure the importance (risk and/or potential negative impact), authority, saturation, and dependency of socio-technical systems to individuals, communities, nation-states or society-at-large in order to signal shifting levels of risk that likely require risk reassessment |
| **Technology Barrier Bias** | Also known as Non-Response Bias, a phenomenon in which the availability, accessibility, and usability of the technology used to gather data or interface with the **AAA System** results in certain participants having reduced ability to participate which affects their representativeness in the dataset, potentially resulting in biassed estimates and discriminatory outcomes |
| **Test Completion Report** | report that provides a summary of the testing that was performed across the end-to-end system including test procedures, metrics, measurements, and thresholds, testing artefacts, and model/system test types.<br><br>[SOURCE: ISO/IEC/IEEE 29119-3:2021, 3.9] |
| **Test Data** | A set of data required to evaluate the Test Plan objectives (e.g., inputs, outputs, ground truth) |
| **Test Item** | work product to be tested |
| **Test Lead** | An expert accountable for the **Test Plan** and **Test Completion Report** for an **AAA System** |
| **Test Plan** | Detailed description of test objectives to be achieved and the means and schedule for achieving them, organised to coordinate testing activities for some test item or set of test items |

| | |
|---|---|
| | [SOURCE: ISO/IEC/IEEE 29119-2:2021, 3.50] |
| **Testing Data** | Data used for providing an independent evaluation of the trained and validated AI system in order to confirm the expected performance of that system before its placing on the market or putting into service |
| **Traceability** | The ability to trace a data right back to its origin through documentation, including a chain-of-custody ("paper trail," physical or otherwise) for data provenance that chronologically records the ownership, viewing, analysis, and transformations of a data record or data sources |
| **Trade-offs** | Decision-making actions that select from various requirements and alternative solutions on the basis of net benefit to the stakeholders [SOURCE: ISO/IEC/IEEE 15288] |
| **Training Data** | Data used for training an AI system through fitting its learnable parameters, including the weights of a neural network |
| **Usability** | A type of user acceptance testing that identifies barriers to usage by considering the needs of impacted stakeholders and **Vulnerable Populations,** including people with a variety of disabilities (e.g., physical, sensory, cognitive, etc.), and assessing the Scope, Nature, Context, and Purpose of the AAA System in terms of use cases, foreseeable scenarios, languages, use of Assistive Technologies, and key modalities of the AAA System |
| **Validation Data** | Data used for providing an evaluation of the trained AI system and for tuning its non-learnable parameters and its learning process, among other things, in order to prevent overfitting |
| **Validity** | The extent to which the results really measure what they are supposed to measure (intended purpose) presently and as time passes distinct from the concept of a (validation) dataset as it relates to training and testing data. |

| | |
|---|---|
| **Version Control and Change Log** | Collects all human deliberative changes, combined with alterations to Pipeline, outcomes, and Architectural Inputs across the lifecycle of the system (Annex IV.6) A description of any change made to the system through its lifecycle; including changes required by a Notified Body |
| **Vulnerable Populations** (People in vulnerable situations) | Persons who often experience exclusion, insufficient accessibility resulting from geopolitical, social, socioeconomic, and cultural inequitable power distribution including but not limited to: children, persons with disabilities, ethnic minorities, and people made vulnerable by an imbalance of power in relation to knowledge, economic or social circumstances, or age |

# 4.0 General Requirements for Accreditation

## 4.1 Interoperability with Standards

ForHumanity's work is designed primarily to ensure an ecosystem called Independent Audit of AI Systems. This ecosystem establishes an infrastructure of trust, predicated on third party, independent assurance of compliance with rules that are either approved by governments and regulators or accepted in the marketplace. This assurance is "at-risk", meaning the independent auditor can be held liable for "false assurance of compliance". As a result of this high standard of behavior, auditors seek maximum, binary clarity on the criteria that determines compliance and non-compliance.

The goal of maximizing the binary (compliant/non-compliant) nature of each audit criteria can be incompatible with industry-led, consensus driven standards from Standards Development Organisations (SDOs). As a result of traditional SDO processes, consensus outcomes sometimes do not reach adequate risk control, treatment, and mitigation for humanity.

Additionally, while some SDOs and their specific standards are accepted widely, some critical national and regional divides occur (such as NIST versus ISO adoption of cybersecurity or risk management in artificial intelligence). This divide makes compliance challenging for corporations acting globally. ForHumanity drafts certification schemes (collection of audit criteria) that are jurisdictionally-sensitive and globally harmonized.

Finally, SDOs are typically industry-led and only recently have begun to factor in a wider perspective of stakeholders. The historical result is that the focus has been on organizational risk management and compliance rather than the risks to the user/AI subject/natural person. ForHumanity's mission is to examine and analyze downside risk associated with the ubiquitous advance of AI, algorithmic and autonomous systems and where possible to engage in risk mitigation to maximize the benefits of these systems… ForHumanity. Therefore, when ForHumanity draft our audit criteria and certification schemes, our different perspective leads us to different human-centric audit criteria.

It is in these three challenges of the SDO process that ForHumanity finds its role. Our primary work is to provide human-centric, binary and globally harmonized audit criteria in support of Independent Auditors and the second-order benefit of facilitating compliance.

As a result of this mission, ForHumanity make the following declarations:

1. In upholding its mission, ForHumanity will ensure that our perspective remains human-centric in our output of audit criteria and certification schemes
2. ForHumanity fully supports the work of SDOs
    a. ForHumanity participates in many SDOs and will continue to expand our efforts to support the development of standards
    b. ForHumanity offers its own crowdsourced, transparent, and expert work in creating binary audit standards that support the development of traditional standards
    c. ForHumanity's audit criteria will always reference accepted, published standards, relevant and consistent with ForHumanity's scope of AI, Algorithmic, and Autonomous (AAA) Systems
3. ForHumanity will ensure that our audit criteria:
    a. Are aligned to accepted, published standards that are legally binding, relevant and consistent with ForHumanity's scope of AI, Algorithmic, and Autonomous (AAA) Systems
    b. Are binary (compliant/non-compliant), implementable, and measurable to accepted forms of evaluation methods for third party independent auditors such as (but not limited to) procedure manuals, published codes, correspondence, physical testing, official filings, pictures/graphics, and contracts
    c. Maximize global harmony, as applicable to facilitate compliance for multi jurisdictional companies

## 4.2 Normative Criteria explanation

Normative criteria take one of three forms shall/should/may and each are described below including how each term is satisfied in the audit certification scheme. All criteria require

documentary evidence, including "may" criterion as they indicate a choice leading to further criteria or disclosures.

**SHALL** - is a requirement. There is no compliance without sufficient satisfaction with the requirements of the criterion. A criterion is a SHALL because it is a legal requirement, a regulatory requirement, or a non-negotiable imperative for the protection of an individual or management/mitigation of a risk to individuals, and has been determined feasible to comply. Strictly from a risk perspective, failure to comply with a SHALL criterion absolutely and unequivocally exposes the organization to risk and non-compliance with the certification scheme.

**SHOULD** - is a recommendation. It is within the power and judgment of an organization to decide if it will comply or not. However, SHOULD identifies the recommended option. Therefore, if the organization makes the choice to not comply, it must recognize and acknowledge that a risk is present and has been accepted. Therefore, audit compliance for a SHOULD statement can take one of two forms. Either documented compliance with the SHOULD statement or documented acceptance of the risk taken, "why" the risk is tolerable, and non-compliance with the criterion is accepted. From a risk perspective, the choice to not comply with a SHOULD statement exposes the organization to risk, but the organization may determine the subsequent risk to be tolerable, unlikely to occur, or mitigated in some other fashion. The assessment and associated mitigations are to be documented.

**MAY** - is a choice without prejudice to the options. It has been determined that compliance or non-compliance with the criterion by itself is neither positive nor negative for humanity inherently. MAY statements will often lead to documented risks that will lead to further compliance requirements based upon the choice. MAY statements exist to clarify for the organization that it does, in fact, have a choice. For audit compliance purposes, the target of evaluation should document the choice it makes. This documentation must also reflect the pros and cons of the choice. Audit compliance is satisfied by this documentation. The choice made in response to a MAY question does NOT mean there is no inherent risk. Each choice has risks associated with it and they should be assessed and documented in the risk assessment process.

## 4.3 Documentation of Assessments and Certification

Certifications may only be conducted by ForHumanity Certified Auditors (FHCA) under contract with accredited entities as established by local accreditation authorities. Certification is available for individuals who demonstrate sufficient knowledge of the scheme and achieve a passing grade on the certification exam.

The following documents shall be produced by the certifying body in order to ensure that the certification is rigorous, transparent, and itself auditable.

- Certification Plan, including:
    - Opening meeting where:
        - The scope is verified
        - Organizations and individuals, including their roles, are documented
    - Confirmation of the authorisation of the Certification Body to award the certification, and their impartiality
    - Description of the ToE (as documented in the contract)
    - Documentation of a process to assess and determine Relevant Legal Framework applicable to the AAA System and associated ecosystem including the role of the Auditee (e.g., Controller/Processor, Provider/Deployer)
    - Expected documentary evidence
    - Physical testing scheduling
    - Any expected deviations from the evaluation methods detailed in the certification criteria
    - Any site or network access required, and any special requirements for that access (e.g. permission to conduct intrusive network scanning)
    - Closing meeting for presentation of Certification Report, issuance of Certification, or issuance of Non-Compliance Letter
- Certification Report that has two versions, a Public version based upon Relevant Legal Framework requirements and a private version for the auditee, including:

    **Public**
    - Public disclaimer including description of the Scope, Nature, Context, and Purpose of the AAA System (Public)
    - The specific dates of inspections (Public)
    - Intended users of the certification report (e.g., investors, clients, regulatory compliance) (Public)
    - Whether a certification is awarded, and its duration (Public)

    **Private**
    - Explanation of the scope, including Beginnings and Ends, agreed in the Audit Engagement Letter (private)
    - Any deviations from the Certification Plan (private)
    - Process narratives, walkthroughs, flowcharts, diagrams, control descriptions, codes, policies (Management Representations) (private, unless required under criteria)
    - The specific software and hardware versions and assets inspected including third-party assets, as applicable (private)

- A list of documentation and assets that will be retained as audit evidence, and explanation of deviations (private)
- A duly authorized signatory (private, but at the auditee's discretion)
- A list of deficiencies if certification will not be issued (private)
- If included in the Audit Engagement Letter, a determination of sufficient/mature levels of compliance (private, but at the auditee's discretion)
- A process for resolving disputes (private)
- A list of non-compliance issues for consideration (private)
- That auditee has met all public disclosure requirements as logged by the auditor (private)
- Sufficient, robust, and resilient ongoing monitoring systems and explicit statement that systemic failures of ongoing monitoring systems will preclude future certification, including next date of expected certification (private)
- Statement of auditor independence and quality management (private)
- Statement of understanding that this certification scheme does not represent complete protection from enforcement of the law by any National Supervisory Authorities (private)

## 4.4 Evaluation Methodology

Each of the scheme criteria identifies an evaluation method type. The certifying body may vary the evaluation method type where it provides additional assurance, but not so that it provides less. The following types are listed:

1. *Contract* - An executed contract can be examined and demonstrates compliance with the criteria.
2. *Correspondence (Internal or External)* - Historical correspondence is available that demonstrates compliance with the criteria.
3. *Employee Handbook* - In the context of an employment contract, an internal document that comprehensively describes an employee's duties, obligations, responsibilities, guidelines, rights, benefits, and available resources.
4. *Internal log, register or database* - Internal, systemic records with proof of authenticity that can be examined by the certifying body and that demonstrate compliance.
5. *Internal procedure manual* - Internal policy and procedure documentation that can be shown to the certifying body to demonstrate compliance with the criteria. These may include, but are not limited to, documents, notifications, interfaces, assessments, studies, rosters, and meeting minutes. All evidence should be of sufficient detail to show that they are up-to-date, implemented, and complete.
6. *Picture/Graphic* - Includes diagrams and technical drawing

7. *Public disclosure document* - Contains all legal obligations and elements as described by the specific audit criteria. The document must meet the definition of Public (as found in Section 3.0).
8. *Physical testing* - At the certifying body's discretion, this can refer to documentation of any of the following:
    a. Interviews with authorized personnel
    b. Inspection of current events, interfaces, and/or notifications
    c. Technical testing including metrics, measurements, and thresholds

Copies of all evidence obtained during the evaluation should be stored in encrypted form by the certifying body, except where the evidence includes personal data and does not comply with the principle of data minimisation.

# 5.0 Criteria catalog

Column 1 = ForHumanity unique identifier (FHUI)
Column 2 = CORE Classification description
Column 3 = Audit criteria
Column 4 = Evaluation method

| FHUI | Categories | Audit Criteria | Evaluation Method |
|---|---|---|---|
| | | Expert Oversight | |
| | Expert Oversight | The **Provider** shall have a duly designated team of experts trained in understanding the following specific and multi-disciplinary risks associated with **AAA System** in regards to:<br>A. Risks to human rights and freedoms of **AI Subjects** (especially **Vulnerable Populations**), including associated legal risks such as equality, nondiscrimination, transparency, and fairness<br>B. Risk of poor data management and | Internal log, register, or database |

| | | | |
|---|---|---|---|
| | | governance, including failure to deliver privacy-by-design and data protection, especially in the areas of sensitive data (e.g., race, gender, age, biometric facial mapping, retinal scan, DNA)<br>C. Risks associated with insufficient risk management processes such as:<br>i. Ineffective risk controls, treatments, and mitigations<br>ii. Ineffective feedback loops<br>iii. Failure to identify incidents<br>iv. Failure to identify and include all stakeholders in risk assessment (need to ensure the inclusion of **Diverse Inputs and Multi Stakeholder Feedback)**<br>D. Risks associated with insufficient disclosure, transparency and operating protocols for the **AAA Systems** including detailing **Residual Risk**<br>E. Risk associated with design and deployment choices regarding machine autonomy, human oversight and interactions<br>F. Risks associated with uncontrolled machine autonomy and inability to assure appropriate human control<br>G. Risks associated with unmitigated **Statistical Bias**, **Cognitive Bias**, and **Technology Barrier Bias**, including managing **Model, Data**, and **Concept Drift**<br>H. Risk of unfair, unreliable, inaccurate and vulnerable model development, including failure to achieve:<br>i. **Causal Hypothesis**,<br>ii. **Construct Validity**,<br>iii. data **Relevancy**,<br>iv. **Data** and **Information Quality** | |

| | | | |
|---|---|---|---|
| | | v. **Accuracy**,<br>vi. Reliability, and<br>vii. Resilience<br>viii. **Ground Truth** validation<br>I. Risk of insufficient oversight of model deployment and ongoing health and fitness for purpose, including inadequate:<br>i. Quality management,<br>ii. Model validation,<br>iii. Post-market and continuous monitoring<br>J. Risk associated with inadequate security and cybersecurity, including risks associated with robust and resilient operations<br><br>(The duly designated team of experts are hereafter referred to as the **Algorithmic Risk Committee** for the purposes of ease of reference.) | |
| | Expert Oversight | The **Provider** shall have a duly designated team of experts trained in understanding the following specific and multi-disciplinary risks associated with **Algorithm Ethics** and **Ethical Choices** associated with **AAA System** such as:<br><br>A. Adjudicating instances of **Ethical Choice,** and analyzing, evaluating and treating Ethical Risk including:<br>i. Human oversight and interactions of the **AAA System** as documented in the **Human Interactions Report**<br>ii. **Controllability**<br>iii. **Necessity**<br>iv. **Explainability** and **Explainability+**<br>B. Compiling the shared moral framework of the organization as applicable to **AAA** | Internal log, register, or database |

| | | | |
|---|---|---|---|
| | | **Systems**, including:<br>i. Identification of applicable **Relevant Legal Frameworks**<br>ii. Documenting the shared moral framework in a **Code of Ethics** and the principles portion of the **Code of Data Ethics**<br>iii. Managing changes to the shard moral framework<br>iv. Establishing an operational definition of diversity, especially for Diverse Input and Multi Stakeholder Feedback<br>C. Managing risks to health, safety and human rights and freedoms of **AI Subjects**, including:<br>i. Assessing and implementing **Proportionality**<br>ii. Assessing **AAA Systems** to determine whether they are:<br>a. Emotional recognition systems<br>b. Social scoring,<br>D. Implementing Fairness in:<br>i. Testing processes, procedures, metrics, measurements, and thresholds<br>ii. **Residual Risk Public** disclosures<br>iii. **Source Data** acquisition<br>E. Evaluating UX/UI interfaces and associated **AAA System** processes and procedures for:<br>i. Detrimental **Nudges**,<br>ii. Deceptive design,<br>iii. Dark patterns<br>iv. Subliminal techniques or other material impacts that distort **AI Subjects'** behavior<br>F. Identifying and mitigating **Cognitive Bias** | |

| | | | |
|---|---|---|---|
| | | in **AAA Systems** anywhere in the algorithmic lifecycle<br>G. Establishing guardrails for **AAA Systems** to identify:<br>i. Deviations from their agreed upon **Scope, Nature, Context, and Purpose,**<br>ii. **Model, Data**, and **Concept Drift**<br>H. Establishing content moderation metrics, measurements, and thresholds, if applicable, to the **AAA System** including:<br>i. **Key Performance Indicators (KPI)** for the risk of **Model Drift, Data Drift**, and **Concept Drift**<br>ii. If applicable, **Key Risk Indicators (KRI)** for **Exceptions Interpretability**<br>iii. If applicable, **Key Detrimental Indicators** for **Content Moderation**<br>iv. If applicable, **Key Regulated Product Indicators** for illegal or harmful product moderation<br>v. If applicable, **Key Language Indicators** for **AAA Systems** using LLM, LMM or questionnaires<br>I. Establishing metrics, measurements and thresholds to control, treat, and mitigate **Ethical Risks** (e.g., vendors, inputs products, services), including the health, safety, and well-being of human interactors<br>J. Assessing systemic riskiness and associated metrics, measurement and thresholds<br>K. Establishing metrics, measurements, and thresholds to identify and classify **Novel technologies** and **Emergent Risk** | |

| | | | |
|---|---|---|---|
| | | ***Note 2 - The duly designated team of experts is hereafter referred to as the Ethics Committee for ease of reference. The controller may refer to this team in any manner they see fit.*** | |
| | | **Top Management and Oversight Bodies** | |
| | Top Management and Oversight Bodies | Top Management and Oversight Bodies shall ensure that the following governance, oversight, and accountability functions for the **AAA System** are operational, including:<br>A. LEADERSHIP AND GOVERNANCE - Demonstrating leadership and commitment to ethical and risk management by establishing standing and empowered **Ethics Committee, Algorithmic Risk Committee,** and all applicable specialty committees (e.g., **Children's Data Oversight, Disability Inclusion & Accessibility,** Digital Services Content**)**<br>B. ACCOUNTABILITY - Delineating roles and responsibilities of the operational teams responsible for organizational-wide compliance and oversight, including a quality management system, legal, compliance, enterprise risk management, and internal audit, and their interactions with **AAA System** specific operations (e.g., the **Algorithmic Risk, Disability Inclusion & Accessibility,** and the **Ethics Committee)**<br>C. RISK MANAGEMENT - Using a currently updated **cAIRE Report**, ensuring that the **Algorithmic Risk Committee** and/or applicable specialty committees are governing and accountable for the risk | Correspondence (Internal or External) |

| | | | |
|---|---|---|---|
| | | management process for the **AAA System**, including:<br>i. Identifying specific and unique risks<br>ii. Ensuring the implementation of risk controls, treatments, and mitigations<br>iii. Monitoring the effectiveness of risk controls, treatments, and mitigations<br>iv. Ensuring that risk management process is conducted over the lifecycle of the **AAA System** and kept current<br>D. REGULATORY COMPLIANCE - Ensuring that the **Algorithmic Risk Committee** works in conjunction with the appropriate **AAA System** expert legal consultation to determine the applicable **Relevant Legal Frameworks** and document a plan for initial and ongoing regulatory compliance for the **AAA System,** including assigning an **AI Compliance Lead**<br>E. RESOURCE ALLOCATION - Ensuring that the necessary resources are allocated to manage governance, oversight, accountability, risk and quality (e.g., people, budget, infrastructure)<br>F. OVERSIGHT - Assigning authority, responsibility, and accountability at appropriate levels within the organization documented in the **Committee Governance Assessment** (conducted by a third line of defense, such as internal audit or enterprise risk management) and ensuring delivery of the assessment to the **Algorithmic Risk Committee**<br>G. DUTY OF CARE FOR VULNERABLE POPULATIONS - Assess stakeholders to identify **Vulnerable Populations**, and then ensure that teams with expertise are established to oversee specific and unique | |

| | | | |
|---|---|---|---|
| | | risks to **Vulnerable Populations** (e.g., **Children**, **Persons with Disabilities**) including the provision of **Accommodations** as applicable.<br>H. STATEMENT OF PRINCIPLES - Endorse:<br>i. A public **Code of Ethics**<br>ii. The principles portion of the **Code of Data Ethics**<br>iii. A commitment to uphold the applicable **Relevant Legal Framework(s)**<br>I. DEFINE STAKEHOLDERS - Ensuring that stakeholders are considered from a holistic perspective of impacted groups, including both direct stakeholders that may be internal (e.g., employees, customers, shareholders) or external (e.g., human users, communities) and indirect stakeholders (e.g., society and the environment)<br>J. DUTY TO STAKEHOLDERS - Ensuring the inclusion of **Diverse Inputs and Multi Stakeholder Feedback** throughout the risk assessment process across the development lifecycle of the **AAA System,** including endorsing the definition of diversity in the **Code of Ethics**<br>K. HUMAN OVERSIGHT - Ensuring that:<br>i. **Humans** are in-the-loop, on-the-loop, **in-Command,** or available for post hoc review of outputs of the **AAA Systems** as appropriate and are duly trained, authorised, and empowered to execute their duties<br>ii. **AAA Systems** always have human ownership and direct legal accountability<br>L. VENDOR MANAGEMENT - Endorsing a policy established by the **Algorithmic Risk** | |

| | | | |
|---|---|---|---|
| | | **Committee** that delineates compliance specifications and liability management requirements for all upstream and downstream suppliers including the **AAA System** itself, data and associated services, networking, storage, and technical infrastructure<br>M. TECHNICAL REQUIREMENTS - Endorsing technical infrastructure investment that is appropriate and proportional to the risk of the **AAA System**, to ensure robust and resilient function that aligns to all **Relevant Legal Frameworks** and industry standards including:<br>i. **Quality Management System,** if applicable<br>ii. Risk Management Framework<br>N. SYSTEM INTEGRITY - Documenting, in a **Code of Data Ethics**, a commitment to robust **AAA Systems** that have<br>i. Data that are:<br>a. **Relevant**<br>b. High quality<br>ii. Architectural inputs that have:<br>a. **Construct validity**<br>iii. Outputs that are<br>a. Validated by **Ground Truth** or<br>b. Disclosed in **Residual Risk** and **Explainability Statements** as inferential conclusions<br>O. TRANSPARENCY - Ensuring the public display of all of the following:<br>i. **Residual Risk**<br>**ii. Data Transparency Document**<br>iii. Public portions of the **Ethical Risk Assessment** | |

| | | | |
|---|---|---|---|
| | | iv. **Explainability Statements** to **AI Subjects**<br>P. TRAINING AND EDUCATION - Ensuring that all employees, including Top Management and Oversight Bodies, and **AI Subjects** are trained and educated, as applicable and according to the **Scope, Nature, Context, and Purpose** of the **AAA System** and the nature of their interactions, on all of the following:<br>i. AI Literacy<br>ii. Risk Management, including **Residual Risk** and potential harms<br>iii. Operating instructions<br>iv. terms and conditions<br>v. Acceptable Use<br>vi. Incident identification and reporting procedures<br>vii. Quality control and assurance standards<br>Q. CHANGE MANAGEMENT - Endorse and communicate a **Change Management Plan** as recommended by the **Algorithmic Risk Committee**<br>R. CONFLICT RESOLUTION - Ensuring that, in all matters where committees (including specialty committees) or delegated persons interact, there is a procedure outlined in the **Code of Ethics** to adjudicate any conflict<br>s. DECOMMISSIONING - Deciding and documenting the decision to decommission the **AAA System**, in consideration of recommendations from the **Algorithmic Risk Committee***[#12]* | |
| | Top Management and Oversight Bodies | Top Management and Oversight Bodies conduct a **Committee Governance Assessment** is that includes the following: | Correspondence (Internal or External) |

| | | | |
|---|---|---|---|
| | | a. Collecting all **Terms of Reference**, reports, logs, assessments, and the **cAIRE Report** with **Traceability** from the **Ethics Committee**, the **Algorithmic Risk Committee,** and any specialty committees (e.g., **Children's Data Oversight Committee**, **Disability Inclusion and Accessibility Committee**)<br>b. Logging all **Duty Designation Letters**, if applicable<br>c. Assessing gaps, inconsistencies, or issues associated with the alignment to the mandates for the specific committees and/or duty designation letters including controls, treatments, and mitigations for identified problems<br>d. Delineating roles and responsibilities between and amongst committees (e.g., **Algorithmic Risk, Ethics Committee**) and leads (e.g., **AAA Cybersecurity, AI Compliance, Quality Management**)<br>e. Identifying all audit criteria that transit from one committee or duly designated officer to another committee or duly designated officer<br>f. Assessing cross communications, sharing of risk inputs, consultations with specific committees including **Ethics Committee,** and/or all specialty committees (e.g., **Children's Data Oversight Committee**) and gaps that exist in such communications or interactions including controls, treatments, and mitigations for identified shortcomings<br>g. Ensuring that user interfaces and **Accommodations Rights Requests**, in regards to the **AAA System**, are integrated, coordinated, and documented with the organization's established accommodation requests process | |

| | | | |
|---|---|---|---|
| | | h. Assessing all committees for:<br>i. Sufficient diversity<br>ii. Sufficient expertise,<br>iii. Conflicts of interest (or duty) to determine disclosure and/or recusal<br>iv. Inclusion of experts (internal or external) from specialty committees onto the **Ethics Committee** and **Algorithmic Risk Committee** for assessments of the specific and unique risks associated with those specialty committees<br>and remediate any shortcomings or conflicts documenting the risk control, treatment, and/or mitigation<br>i. Recording all risk controls, treatments, mitigations, and **Residual Risk** in the **cAIRE report**<br>j. Endorsing the accepted **Residual Risk**<br>k. Endorsing the **Decommissioning Policy,** process and procedure | |
| RRRR | Top Management and Oversight Bodies | In consideration of:<br>1. The **cAIRE Report**, including:<br>i. Current **Residual Risk**<br>ii. The advised metrics, measurements and thresholds applicable to **Risk Appetite** and **Risk Tolerance** associated with the **AAA System**<br>2. Organizational **Risk Appetite** and **Risk Tolerance**<br>Top Management and Oversight Bodies shall execute the following steps to establish the accepted **Residual Risk**:<br>A. Assess to determine whether additional risk controls, treatments, and mitigations for the **AAA System** are to be implemented<br>B. Assess to determine whether external risk treatment options are to be implemented | Correspondence (Internal or External) |

| | | | |
|---|---|---|---|
| | | C. Assess to determine the current **Residual Risk** is accepted with **Traceability** | |
| | Top Management and Oversight Bodies | The Top Management and Oversight Bodies shall ensure, with **Traceability**, that all logs, records, and assessments related to risk (e.g., **Algorithmic Risk Assessment**, **Ethical Risk Assessment**) are documented in the **cAIRE Report** and provided to enterprise or organizational risk management logs, registers, or databases | Correspondence (Internal or External) |
| | Top Management and Oversight Bodies | Top Management and Oversight Bodies shall ensure that a person educated on **Ethical Choice** and **Algorithm Ethics,** or equivalent, from the **Ethics Committee** duly designated to assist the **Algorithmic Risk Committee** in managing the risks from **AAA System***[#20]* | Internal procedure manual |
| Relevant Legal Framework and Modular Assurance Assessments | | | |
| | Relevant Legal Framework | The **Algorithmic Risk Committee** shall assess the deployment of the **AAA System** to determine all applicable **Jurisdictions** in which the **AAA System** is deployed and document those **Jurisdictions** in the **AAA Systems List** | Internal log, register, or database |
| | Relevant Legal Framework | For each **Jurisdiction** in which the **ToE** operates and in the context of its **Scope, Nature, Context,** and **Purpose,** and in consultation with the legal team (internal and/or external), the **Algorithmic Risk Committee,** shall regularly or as needed (e.g., as laws or jurisprudence change) assess the **AAA System** for applicable legal obligations including, but not limited to, the following sectors of law: | Internal log, register, or database |

| | | | |
|---|---|---|---|
| | | A. Fundamental and Human Rights<br>B. Legal/Lawful basis<br>C. Data collection, protection and retention<br>D. Equality and nondiscrimination<br>E. Access to goods and services<br>F. Market and competition law<br>G. National Security<br>H. Prohibited Systems<br>I. Sector-specific law (e.g., health, security)<br>J. Protection for Vulnerable Populations (e.g., Elderly, Children, Persons with Disabilities)<br>K. Employment law<br><br>and document the following details in the **Relevant Legal Framework** log:<br>1. Applicable legal obligations as **Relevant Legal Frameworks**<br>2. A conclusion, from the legal expert, that the ToE is compliant with applicable legal obligations, prior to placing the **AAA System** on market<br>3. Legal expertise of the person providing the legal opinion, including certification from oversight bodies where applicable | |
| | Modular Assurance Assessment | If the **Provider** deploys the **AAA System** directly to **AI Subjects**, then the **Provider** shall ensure that all **Deployer** obligations are met by documenting separate assurance under the ForHumanity CORE Deployer Certification scheme or equivalent | Internal log, register, or database |
| | Modular Assurance Assessment | In consideration of:<br>1. **Relevant Legal Frameworks**<br>2. **Jurisdictions** of operation,<br>The **Algorithmic Risk Committee** shall assess the **ToE** to determine whether **Personal Data** is | Internal log, register, or database |

| | | | |
|---|---|---|---|
| | | processed and document the conclusion in the **AAA Systems List** | |
| | Modular Assurance Assessment | In consideration of:<br>1. **Relevant Legal Frameworks**<br>2. **Jurisdictions** of operation,<br>and in consultation with the **Ethics Committee** and legal experts (internal or external), the **Algorithmic Risk Committee** shall assess the **AAA System** to determine if **Children** are **AI Subjects** and document the conclusion in the **AAA Systems List** | Internal log, register, or database |
| | Modular Assurance Assessment | In consideration of:<br>1. **Relevant Legal Frameworks**<br>2. **Jurisdictions** of operation,<br>And in consultation with the **AAA Cybersecurity Lead** and legal experts (internal or external), the **Algorithmic Risk Committee** shall assess the **AAA System** to determine whether there are legal obligations pertaining to security and/or cybersecurity and document the conclusion in the **AAA Systems List** | Internal log, register, or database |
| | Modular Assurance Assessment | In consideration of:<br>1. **Relevant Legal Frameworks**<br>2. **Jurisdictions** of operation,<br>And in consultation with the **Ethics Committee** and legal experts (internal or external), the **Algorithmic Risk Committee** shall assess the **AAA System** to determine whether there are applicable use case or industry specific:<br>A. Legal obligations<br>B. Standards (e.g., harmonized standards, common specifications)<br>C. Voluntary Standards (e.g., ForHumanity certification schemes)<br>and document the conclusion in the **AAA Systems List** | Internal log, register, or database |

| | | | |
|---|---|---|---|
| | | | |



# Appendix A - Infrastructure of Trust for AI - Guide to Entity Roles and Responsibilities

ForHumanity promotes this certification scheme to all entities that wish to provide advice, guidance, consulting and assurance or organisation both outside and within the European Union. ForHumanity licences entities to offer the scheme and a list of licensed entities can be found here.

ForHumanity also trains individuals to become ForHumanity Certified Auditors (FHCAs). Earning this certification is the ultimate assurance of knowledge of this certification scheme and the process by which certification is achieved for organisations. ForHumanity offers online, asynchronous training in this certification scheme through its training platform – ForHumanity University.

ForHumanity promotes this and many other certification schemes to organisations, governments, regulators, national accreditation bodies, professionals and the public by social media, conference speeches, university lectures, online presence, and execution of our mission statement to specific support an infrastructure of trust with a wide range of participants from society.

Describing the roles in an infrastructure of trust for AI, Algorithmic and Autonomous (AAA) Systems - we have a model with a long track record of success. ForHumanity is adapting that model to AAA Systems.

## Background on Independent Audit

In 1973, the accounting industry came together and formed The Financial Accounting Standards Board (FASB) which created the Generally Accepted Accounting Principles (GAAP) which still govern financial accounting today. Eventually, the US Securities and Exchange Commission, and other extranational regulatory agencies, required adherence to the GAAP standard for all publicly listed companies. This clarity and uniformity significantly improved the financial world. An infrastructure of trust has been built over the past 50 years because of critical features such as independence, certified practitioners, and third-party rules that are compliant with the law and best practices.

## Adapting to AI and Autonomous Systems

ForHumanity has advocated for the adoption of this infrastructure of trust and explained how it can be adapted and adopted for the Governance, Accountability, and Oversight of AI and Autonomous Systems. We support the creation and mandate of Independent Audit of AI Systems (IAAIS). IAAIS provides a comprehensive solution grounded in the same fundamental principles as Independent Financial Audit. ForHumanity develops and maintains audit and certification criteria designed for a range of industries and jurisdictions.

The proposed system replicates the distributed oversight, accountability and governance needed for AI, Algorithmic, and Autonomous (AAA) systems in the same manner as financial audit, through audit and pre-audit service providers. These entities will employ certified practitioners to prepare for an eventual

independent audit performed by other certified practitioners. The audit criteria are crowdsourced and presented transparently to maximise an entity's ability to achieve compliance. Advancements in systems technology allow many of these processes to be automated for entities such as with the Treadway Commissions' Committee of Sponsoring Organization (COSO) framework for internal risk, audit and controls. The result is a fully-integrated, compliance-by-design infrastructure that embeds human agency, transparency, disclosure and compliance from design to decommission.

## Role on Independent Audit of AI and Autonomous Systems

The audit criteria are applied in two vectors: 1) Top-down accountability, governance and oversight 2) laterally, AI system by AI system. The top-down approach creates accountability systems for ethics, bias, privacy, trust, and cybersecurity for the Board of Directors, Chief Executive Officer, and Chief Data Officer. Committee structures are required such as an Algorithmic Risk, Ethics and specialty committees to manage the audit/compliance responsibilities, as a second line of defence. All of these top-down criteria apply to every AI and every autonomous system in the organisation. The system-specific audit criteria are designed to ensure legal and best practice compliance tailored to the specific impact of each system on humans. This comprehensive approach ensures consistency across the organisation combined with complete risk management coverage of each unique system.

## Participants in the System

The roles largely remain the same in Independent Audit of AI Systems as described in Taxonomy. There are six distinct roles in most jurisdictions. Each player performs their function and the rules are executed in the same conflict-free manner, ensuring the highest integrity.

**Certifying Bodies/Notified Bodies/Auditors (Auditors)**

- An Auditor engages in 3-party contract party contracts, with the Target of Evaluation (ToE) and on behalf of the public or intended users.
- The auditor deploys certified practitioners to conduct the audits.
- The auditor itself is certified by the Government Accreditation Service.
- When audits are conducted there is no feedback loop to the company and the audit is compliant or non-compliant.
- Audits are publicly disclosed according to the rules of the jurisdiction.
- The Auditor is liable for false assertions of compliance
- An Auditor is licensed for use of certification criteria
- The Auditor shall not provide Pre-audit services to Audit clients
- An Auditor may provide Pre-Audit services to non-Audit ToEs (may require accreditation)

**Pre-Audit Service Providers/Consultants/Advisors (PASP)**

- PASP engages in a 2-party contract directly with the Target of Evaluation
- There is a direct feedback loop between the ToE and PASP
- The PASP may or may not deploy certified practitioners per local jurisdiction rules
- The PASP may or may not be accredited by the Government Accreditation Service
- The PASP offers no certification or guarantee of audit compliance
- The PASP works are private, on behalf of the ToE

- The PASP is not liable for failed compliance or false assertions of compliance
- The PASP may or may not be licensed for use of certification criteria, but must be licensed if the service offered is related to or designed to satisfy certification requirements
- The PASP shall not be the auditor for a PASP client
- A PASP may offer Audit service to non-PASP clients (must be accredited)
- A PASP may deploy compliance-in-a-box solutions for criteria compliance

**Entities seeking Certification/Providers/Deployers (Auditee)**

- Auditee may engage PASP
- Auditee shall have an Auditor if required by the Relevant Legal Framework
- Auditee pledges that all components, systems and relevant, supporting infrastructure to be certified will be disclosed to the Auditor, failure in this regard is the responsibility of the ToE
- Auditee dealings with PASP shall be confidential and non-public audit compliance may be confidential with an Auditor
- Auditee shall maintain compliance structures, such as Algorithmic Risk Committee, Children's Data Oversight Committee, and Ethics Committee
- Auditee shall build and maintain internal controls and systems to aid in compliance with audit requirements and foster robust risk management, monitoring, and regulatory compliance
- Auditee shall be responsible for all public disclosures

**Third-Party Criteria creation, maintenance, and individual certifier (ForHumanity)**

- Non-profit organisation
- Independent of Auditors and PASP
- Transparent and inclusive of input and critique from all participants
- Criteria designed to uphold human well-being
- Conflict-free of undue Auditee influence
- Submits to the authority of the jurisdiction for certified criteria
- Iterates and maintains criteria consistent with the law and best practices in a binary and auditable fashion
- Trains and certifies individual practitioners on all criteria in support of uniformity of audit assurance process
- Maintains a transparent repository of use cases and knowledge stores in support of Auditors/Auditees to facilitate compliance
- Licences criteria to all qualified Certifying Bodies/Notified Bodies/Auditors/PASP
- Provides standard contract clauses for Auditors and PASP
- Engages in distributed education system to maximise availability and certified individuals
- Maintains a system of Continuing Education (CE)
- Maintains a searchable, registration system of Accredited Individuals and holds them to a Code of Ethics and Professional Conduct
- Ensures Independence and anti-collusion amongst of Certifying Bodies/Notified Bodies/Auditors/PASP
- Maximises global harmony amongst audit criteria while ensuring jurisdictional sensitivity

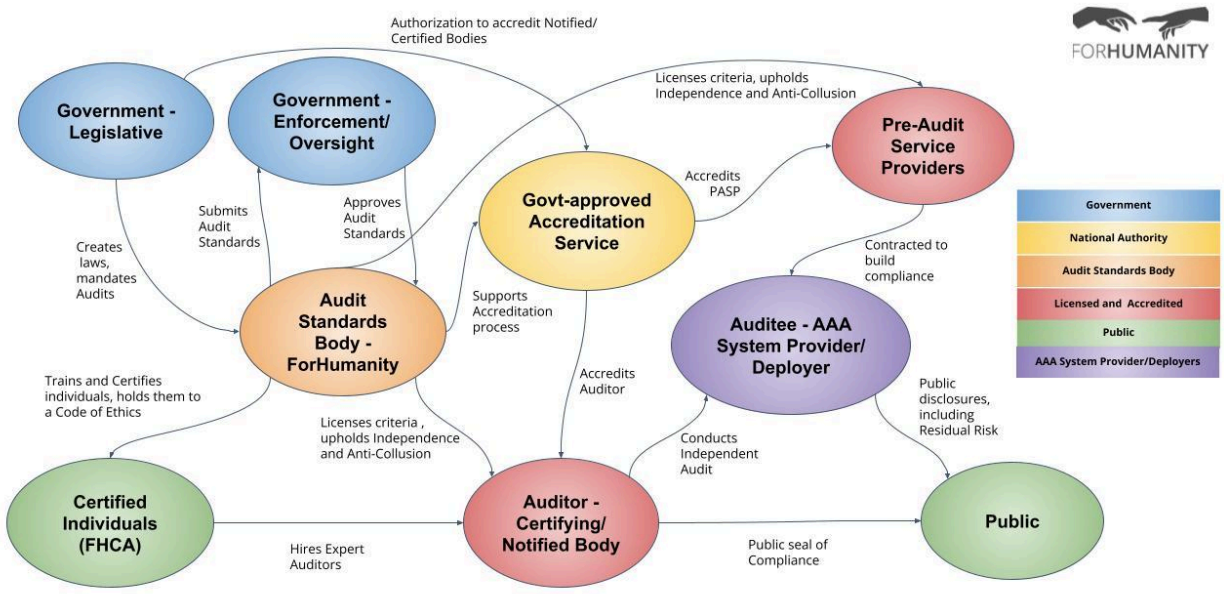
**Government-approved Accreditation Service**

- Creates trust and confidence in products and services
- Assures that Certified/Notified/Accredited Bodies have sufficient talent, skill, scope, and financial foundation to provide certification
- Regular review of accreditation standards
- Regular review of Certified/Notified/Accredited Bodies
- Regular review of Third-party Criteria provider and individual certification
- Determines form and elements of Post Audit Compliance Report
- Maintains an accessible list of Certified/Notified/Accredited Bodies
- Maintains an accessible list of sanctioned or suspended Certified/Notified/Accredited Bodies

**Governments/Regulators or similar Law-making/enforcement body**

- Democratically, elected body
- Legislative responsibilities
- Executive or enforcement responsibilities
- Establishes prohibited AAA Systems
- Establishes low risk and exclusionary criteria from mandatory Independent Audit
- Regularly meets to review laws and best-practices
- Establishes a panel of experts to reviews and accredits (or rejects) submitted criteria
- Engages in enforcement actions for non-compliance with the law
- Handles concerns and issues brought by the Public




**Roles and Responsibilities - Independent Audit - Infrastructure of Trust**

# Licensing

ForHumanity provides four types of licences:

A. Auditor/Certification Body and Pre-Audit Service Provider
B. Platform, technology, or SaaS tools
C. Teaching (for commercial purposes)
D. University (for academic and research purposes) as well as commercial use of certification course

Any entity that uses the certification scheme as the basis of their business relationship (generating revenue or a similar quid pro quo - commercial purposes) with a client must be duly licensed. Any organisation may be licensed by ForHumanity, but they must also have FHCAs on staff in good standing to issue certificates or provide services using the intellectual property.

Audit fees are owed upon receipt of revenue by a licensee. The licence fees allow ForHumanity to maintain the certification schemes and training individuals as experts or ForHumanity Certified Auditors (FHCA). Trademarks, certification marks, audit criteria, and services marks of ForHumanity are provided in licensing agreements and must be used in adherence with the terms of service found in the licence agreement. All licence agreements contain identical terms and conditions as relatable across use cases and are non-negotiable to ensure uniformity.