Responsible AI Certification Assessment Criteria

I. INTRODUCTION 2

II. ASSESSMENT CRITERIA 3

VALID AND RELIABLE 3

EXPLAINABLE AND INTERPRETABLE 6

ACCOUNTABLE AND TRANSPARENT 8

PRIVACY-ENHANCED 13

FAIR – WITH RISK OF HARMFUL BIAS MANAGED 15

SAFE 19

SECURE AND RESILIENT 22

III. DEFINITIONS 24

I. INTRODUCTION

The TrustArc Responsible AI Program is designed to assess and, where eligible, certify the practices of organizations that deploy AI Systems against these Assessment Criteria, which aim to aid certified organizations in demonstrating that they have responsible AI System deployment practices in place.The Responsible AI Program uses a risk-based approach ensuring the obligations associated with the deployment of AI Systems are aligned with the level of risk it entails.

The Assessment Criteria for this program take into account multiple AI frameworks including, but not limited to, OECD AI Principles, NIST AI Risk Management Framework, ISO/IEC 42001:2023, The White House Blueprint for an AI Bill of Rights, NCSC Guidelines for Secure AI System Development, and the EU AI Act, as well as TrustArc's Nymity Privacy Management Accountability and Privacy and Data Governance Framework(s).

The Assessment Criteria are organized into seven sections:

- Valid and Reliable
- Explainable and Interpretable
- Accountable and Transparent
- Privacy-Enhanced
- Fair - with Risk of Harmful Bias Managed
- Safe
- Secure and Resilient

Each section contains relevant Assessment Criteria used to assess an organization's compliance with the Responsible AI Program. Mapping of the Assessment Criteria to the TrustArc Framework standards and controls and external regulatory standards are noted next to the Assessment Criteria.

Any organization participating in a TRUSTe Assurance Program or "Program" agrees to comply with TRUSTe's Assurance Program Governance Standards, which apply to all Programs, and the Assessment Criteria of any Program in which the organization chooses to participate. The Assurance Program Governance Standards ensure that the Program is meaningful and effective in its implementation of robust mechanisms to:

- review and enable organizational demonstration of compliance with the Assessment Criteria;
- enable individuals to raise concerns about a participating company's compliance with the Assessment Criteria via TRUSTe's independent dispute resolution mechanism; and
- address any actual or perceived non-compliance from a participating company against the Assessment Criteria, which includes being able to take actions, including, but not limited to revocation of the company's certification or verification, and any associated seals.

Upon successful completion of the TRUSTe assessment and certification processes, organizations participating in this Program will be issued an attestation letter and authorized to display the TRUSTe Responsible AI Certified seal.

Defined terms appear in **bold.**

II. ASSESSMENT CRITERIA

VALID AND RELIABLE

Validity and reliability are key to ensuring the trustworthiness of an **AI System**. Regularly monitor and audit deployed **AI Systems** to confirm that they are performing as intended and that outputs are accurate and reliable, and include oversight by human operators.

TrustArc P&DG Framework and External Regulatory Standard Mapping	Assessment Criteria
TrustArc P&DG Control *Policies and Standards 1.5:* Develop policies, procedures, and guidelines to define and deploy effective and sustainable governance and controls for managing data-related risks. **TrustArc P&DG Control** *Monitoring and Assurance 3.1:* Continually monitor and periodically evaluate program maturity, and periodically assess and audit the effectiveness of program controls and risk-mitigation initiatives. EU AI Act Article 29 Obligations of deployers of high-risk AI systems ISO/IEC 42001 9.1 Monitoring, Measurement, Analysis and Evaluation ISO/IEC 42001 B.9.4 Intended use of the AI system NIST AI RMF GOVERN 2.1 NIST AI RMF MEASURE 2.5	**1. Monitor Intended Use** Requirement: The **Participant** must monitor the **AI System** to ensure that it is performing as intended during the tenure of the deployment to the extent that it is technically feasible and commercially reasonable to monitor system performance, and that outputs are consistent with the **Participant's** expectations. Monitoring should consider whether the **AI System** is performing according to the intended use as described by the AI developer and according to the use intended by the **Participant**. Mechanisms may include policies or procedures that address the following: ● who will be involved in monitoring and analysis ● what needs to be monitored ● the methods for monitoring, analysis and evaluation to validate (e.g., spot checks, output validation) ● the frequency of the monitoring ● performance according to intended use ● when monitoring and analysis will be performed ● defined relevant monitoring metrics. The **Participan**t must have a level of knowledge to ensure it is capable of understanding the **AI System's** outputs. If an **AI System** does not perform according to its intended use, the **Participant** must update and retrain the **AI System's** inputs and/or outputs, to the extent the **Participant** has control over the inputs, so that the **AI System** performs according to its intended use, or have a suitable compensating control (e.g., a disclaimer or terms and conditions regarding what can or cannot be placed into the **AI System**). Otherwise, taking into account the purpose of processing and where it is likely to result in decisions that produce adverse legal or similarly significant effects,

	including, but not limited to, any risk to rights and freedoms of **individuals**, the **Participant** may need to cease use of the **AI System** and inform the AI developer if the system does not perform to its intended use. Evaluation: TRUSTe will verify that the **Participant** monitors deployed **AI Systems** and routinely assesses that **AI Systems** perform according to the intended use as described by the AI developer to the extent the Participant is able to monitor system performance, and according to the use intended by the **Participant**. Gaps and Remediation: If the **Participant** indicates it does not monitor **AI Systems** for intended use, TRUSTe must inform the **Participant** that the existence of mechanisms (e.g., policies and procedures) to monitor **AI Systems**, where available and commercially reasonable, for intended use is required for compliance with this requirement and/or put in place a suitable compensating control.
TrustArc P&DG Control *Monitoring and Assurance 3.1:* Continually monitor and periodically evaluate program maturity, and periodically assess and audit the effectiveness of program controls and risk-mitigation initiatives. NIST AI RMF GOVERN 3.2 NIST AI RMF MEASURE 2.3, 2.4, and 2.5	**2. Audit Accuracy of Outputs** Requirement: The **Participant** must audit the accuracy of the **AI System's** outputs if the output is likely to result in decisions that produce adverse legal or similarly significant effects, which may include, but not limited to, any risk to rights and freedoms of **individuals**. Mechanisms may include QA testing procedures to measure accuracy that address the following: ● a description of the methodology used to measure accuracy/how will accuracy of AI outputs be monitored after the AI is deployed ● clearly defined and realistic test sets that are representative of the conditions of intended use ● false positive and false negative rates ● human involvement in AI decision-making. Mechanisms (e.g.,QA testing procedures) should be in place to assess the accuracy and quality of generated outputs when new inputs are added. Evaluation: Using the above, TRUSTe will verify that the **Participant** audits the accuracy of the **AI System's** outputs if the output is likely to result in decisions that produce adverse legal or similarly significant effects, which may include, but not limited to, any risk to rights and freedoms of **individuals.** Gaps and Remediation: If the **Participant** indicates it does not audit the accuracy of outputs from the **AI Systems** it deploys, then TRUSTe must inform the **Participant** that the existence of audit mechanisms is required to comply with this requirement if the output is likely to result in decisions that produce adverse legal or similarly significant effects, which may include, but not limited to, any risk to rights and freedoms of **individuals**.

TrustArc P&DG Control *Processes 1.9.2:* Establish, manage, measure, and continually improve processes for implementing all necessary controls to mitigate risk to appropriate levels. ISO/IEC 42001 B.9.3 Objectives for responsible use of AI system NIST AI RMF GOVERN 3.2 and 3.5 NIST AI RMF MAP 3.5 NIST AI RMF MEASURE 2.4	**3. Human Oversight** Requirement: Taking into account the purpose of processing and where it is likely to result in decisions that produce adverse legal or similarly significant effects, which may include, but not limited to, any risk to rights and freedoms of **individuals**, the **Participant** must have mechanisms in place for human oversight where the **AI System** ● cannot detect or correct its own errors, or ● requires human oversight for the **AI System** to function properly according to its instructions or other documentation for its intended use. Mechanisms for human oversight, where required, are defined, assessed, and documented in accordance with organizational policies, Human operators must have the authority to override errors and decisions made by the **AI System**. Mechanisms may include policies and procedures that address: ● how a human operator is notified of/becomes aware when inaccurate outcomes occur (e.g., proactive checks, monitors outputs, random audits) ● the procedure for how a human operator corrects errors in the **AI System** Evaluation: TRUSTe will verify that the **Participant** has mechanisms in place for human oversight where the **AI System** cannot detect or correct errors, or if required for acceptable use of the **AI System** according to instructions or other documentation associated with the intended deployment of the **AI System**. Gaps and Remediation: If the **Participant** does not have human oversight mechanisms in place, where they are required, TRUSTe must inform the **Participan**t that the existence of human oversight mechanisms is required for compliance with this requirement. Where the **Participant** identifies that the **AI System** self-detects and corrects errors, or is not required to comply with the developer's instructions for acceptable use, TRUSTe must verify whether this is the case.

EXPLAINABLE AND INTERPRETABLE

Ensure there is documentation about how the **AI System** works, and that its reviewed and updated regularly. Describe how and why a decision is made in the **AI System**, and its meaning. Tailor descriptions and explanations to the target audience.

TrustArc P&DG Framework and External Regulatory Standard Mapping	Assessment Criteria
TrustArc P&DG Control *Policies and Standards 1.5:* Develop policies, procedures, and guidelines to define and deploy effective and sustainable governance and controls for managing data-related risks. **TrustArc P&DG Control** *Monitoring and Assurance 3.1:* Continually monitor and periodically evaluate program maturity, and periodically assess and audit the effectiveness of program controls and risk-mitigation initiatives. ISO/IEC 42001 7.5.1 Documented information - General NIST AI RMF MAP 2.2 and 3.5 NIST AI RMF MEASURE 2.9	**4. Document System Decisions** Requirement: The **Participant** must document or have documentation about how the **AI System** works and how it makes decisions/generates outputs. Information must be provided about the mechanisms underlying the **AI System's** operation, the **AI System's** knowledge limits, and where used for automated decision making, identify how system outputs may be used. Documentation should provide sufficient information to assist relevant **Personnel** when making decisions and taking subsequent actions. Evaluation: TRUSTe must verify the existence of documentation that represents how the **AI System** works and makes decisions/generates outputs. Gaps and Remediation: If the **Participant** does not have documentation about how the **AI System** makes decisions and the intended use of those decisions, TRUSTe must inform the **Participant** that documentation of this nature is required for compliance with this requirement.
TrustArc P&DG Control *Policies and Standards 1.5:* Develop policies, procedures, and guidelines to define and deploy effective and sustainable governance and controls for managing data-related risks.	**5. Review and Update Documentation** Requirement: The **Participant** must regularly review documentation explaining how the **AI System** works and makes decisions/generates outputs, and update or request updated documentation where necessary. Policies and procedures should be in place that address the following: ● the frequency of reviews ● date and time tracking of reviews

TrustArc P&DG Control *Policies and Standards 1.5.2:* Document and communicate updates to policies, procedures, and guidelines. ISO/IEC 42001 7.5.2 Creating and updating documented information	● the personnel involved in the review ● what decisions/outputs are reviewed ● what updates are made, if necessary. Evaluation: TRUSTe must verify the existence of policies and procedures to ensure the regular review of explanatory documentation. If the **Participant** is using an AI System purchased off-the-shelf from a third party, TRUSTe will verify whether the Participant is able to obtain updated documentation from the third party if the output is likely to result in decisions that produce adverse legal or similarly significant effects, which may include, but not limited to, any risk to rights and freedoms of **individuals**. Gaps and Remediation: If the **Participant** does not regularly review explanatory documentation, TRUSTe must inform the **Participant** that it must put in place policies and procedures to ensure explanatory documentation is regularly updated unless it is a third party off-the-shelf **AI System** that is not used for automated decision making or the outputs do not produce adverse legal or similar significant effects, which may include, but not limited to, any risk to rights and freedoms of **individuals.**
TrustArc P&DG Control *Policies and Standards 1.5:* Develop policies, procedures, and guidelines to define and deploy effective and sustainable governance and controls for managing data-related risks. **TrustArc P&DG Control** *Reporting and Certification 3.2* Select and implement mechanisms to demonstrate the effectiveness of your program and controls to management, the Board of Directors, employees, customers, regulators, and the public. ISO/IEC 42001 B.7.2 Data for development and enhancement of AI system NIST AI RMF GOVERN 3.2 NIST AI RMF MEASURE 2.9 and 3.1	**6. Accessible Information and Explanations** Requirement: The **Participan**t must make accessible information and explanations that will help **Personnel** and vendors understand how the **AI System** works and why the **AI System** makes decisions/generates outputs. The **Participant** should be able to describe how data is used to determine an **AI System's** output. Descriptions should be appropriate to the stage of the AI lifecycle, and tailored based on risk level and to individual differences in the target audience, such as their role, knowledge, and skill level. Employ a variety of methods to explain **AI Systems**, such as visualizations, model extraction, and feature importance. Evaluation: TRUSTe will verify that the **Participant** provides access to information to help **Personnel** and vendors interacting with the **AI System** understand how the system functions, and why it generated a particular output. Gaps and Remediation: If the **Participant** indicates that it does not make explanatory information accessible to **Personnel** and vendors, TRUSTe must inform the **Participant** that it must make accessible information and explanations that will help them understand how the **AI System** works, and how certain decisions are made and/or outputs are generated.

OECD AI Principle 1.3 Transparency and explainability	

ACCOUNTABLE AND TRANSPARENT

Ensure the role and impacts of data used in **AI Systems** are understood throughout the **AI System's** life cycle. Inform **Individuals** when they are interacting with AI, the ways in which data about them are processed, and their rights related to AI data.

TrustArc P&DG Framework and External Regulatory Standard Mapping	Assessment Criteria
TrustArc P&DG Control *Policies and Standards 1.5:* Develop policies, procedures, and guidelines to define and deploy effective and sustainable governance and controls for managing data-related risks. ISO/IEC 42001 A2.2 AI policy NIST AI RMF GOVERN 1.2 NIST AI RMF MAP 3.4	**7. AI Policy** Requirement: The **Participant** must establish an AI Policy that describes the **Participant's** use and other acceptable uses of AI, and establish mechanisms to standardize and communicate the policy with **Personnel** and vendors interacting with the **AI System**. The AI Policy or suitable alternatives should substantially address the use of, restrictions to, and risk management around **AI Systems.** Examples of the preceding may include the following concepts: ● key terms and concepts related to **AI Systems** and the scope of their purposes and intended uses ● the need for business justification for using AI ● acceptable uses of AI, including the acceptable amount of drift from baseline performance ● expected and potential risks and impacts ● an overarching vision for AI usage and growth in the organization, including mission statements, clear objectives, and/or KPIs that align with this vision ● detailed information about regional, industry-specific, and relevant regulatory compliance laws as well as other ethical considerations ● a catalog of approved tools and services that can be used for AI deployment purposes ● clearly defined roles and responsibilities related to the usage and management of **AI Systems** ● data privacy and security mechanisms ● defined procedures for reporting and addressing AI performance and security issues ● standards for AI model performance evaluation (i.e. adversarial testing / red-teaming)

	● consequences and outcomes for violations of the policy. Accountability practices should be proportional and proactive to the severity of consequences posed by the use of the **AI System**. The **Participant** may communicate the AI Policy through mechanisms such as: ● email ● employee intranet ● vendor contracts and/or agreements Evaluation: TRUSTe must verify that the **Participant**: ● has a documented AI Policy or similar policies that addresses, at a minimum, the uses of, restrictions to, and risk management around AI within the organization; and ● communicates the applicable policy with relevant **Personnel** and vendors who engage with the **AI System**. Gaps and Remediation: If the **Participant** does not have a documented AI Policy or similar policies and/or has not communicated it to relevant **Personnel** and vendors who engage with the **AI System**, TRUSTe must inform the **Participant** that a documented AI Policy or similar policy describing, at a minimum, the uses of, restrictions to, and risk management around AI within the organization and its subsequent communication with relevant **Personnel** and vendors is required for compliance with this requirement.
TrustArc P&DG Control *Policies and Standards 1.5:* Develop policies, procedures, and guidelines to define and deploy effective and sustainable governance and controls for managing data-related risks. ISO/IEC 42001 A.7.5 Data Provenance ISO/IEC 42001 B.4.3 Data Resources ISO/IEC 42001 B.7.5 Data Provenance NIST AI RMF MAP 3.4	**8. Data Governance** Requirement: The **Participant** must identify and document information about the data used for the **AI System**, such as through a Data Governance Policy, or similar documentation [e.g., record of processing activity (ROPA), data inventory, or a data protection impact assessment (DPIA)]. Documentation should include but is not limited to, the following topics: ● the provenance of the data used in **AI Systems** over the life cycles of the data and the **AI System**, including the categories of data collected (proprietary information, personal and non-personal), the origins of data, and the original purpose of collection (if **Personal Information** or proprietary information) ● the date that the data were last updated or modified (e.g., date tag in metadata) ● defined roles for AI governance including the person or team responsible ● for machine learning, the categories of data (e.g., training, validation, test, and production data) ● categories of data ● process for labeling data ● intended use of the data ● quality of data ● applicable data retention and disposal policies

	● known or potential bias issues in the data ● the criteria for selecting data preparations and the data preparation methods to be used. <u>Evaluation:</u> TRUSTe must verify the existence of a Data Governance Policy or similar documentation to ensure that the handling of data used in the **AI System** is appropriately being tracked and managed. <u>Gaps and Remediation:</u> If the **Participant** does not have a Data Governance Policy or similar documentation, TRUSTe must inform the **Participant** that a Data Governance Policy or similar documentation is required for compliance with this requirement.
TrustArc P&DG Control *Resource Allocation FC1.4:* Allocate appropriate resources to support the defined mission and vision, and to manage identified risks. **TrustArc P&DG Control** *Policies and Standards FC1.5:* Develop policies, procedures, and guidelines to define and deploy effective and sustainable governance and controls for managing data-related risks. **TrustArc P&DG Control** *Processes 1.9.2:* Establish, manage, measure, and continually improve processes for implementing all necessary controls to mitigate risk to appropriate levels. **TrustArc P&DG Control** *Reporting and Certification 3.2* Select and implement mechanisms to demonstrate the effectiveness of your program and controls to management, the Board of Directors, employees, customers, regulators, and the public.	**9. Alert on Adverse Outcomes** <u>Requirement:</u> The **Participant** must have mechanisms in place to alert human operators to adverse outcomes or impacts of the **AI System** if the output is likely to result in decisions that produce adverse legal or similarly significant effects (e.g., material security or reputational concerns), which may include, but not limited to, any risk to rights and freedoms of **individuals.** Mechanisms may include regular auditing, procedures to monitor the **AI System's** operation for reported issues and failures, and capabilities for external parties to report adverse impacts (e.g., unfairness). Transparency should encompass human-AI interaction: for example, how a human operator or others interacting with the system are notified when a potential or actual adverse outcome caused by an **AI System** is detected. <u>Evaluation:</u> TRUSTe must verify the existence of regular auditing or mechanisms, such as policies and procedures, to ensure human operators are alerted to adverse outcomes or impacts of the **AI System** if the output is likely to result in decisions that produce adverse legal or similarly significant effects (e.g., material security or reputational concerns), which may include, but not limited to, any risk to rights and freedoms of **individuals**. <u>Gaps and Remediation:</u> If the **Participant** does not have in place mechanisms to alert human operators of adverse outcomes or impacts or regular auditing, TRUSTe must inform the **Participant** that such mechanisms are required for compliance with this requirement if the output is likely to result in decisions that produce adverse legal or similarly significant effects (e.g., material security or reputational concerns), which may include, but not limited to, any risk to rights and freedoms of **individuals**.

ISO/IEC 42001 B.8.3 External reporting NIST AI RMF MEASURE 3.1	
TrustArc P&DG Control *Disclosure to Third Parties and Onward Transfer 2.7:* Execute appropriate contracts with vendors supporting the process or technology or with any third parties. NIST AI RMF GOVERN 6.1 TRUSTe Enterprise Privacy & Data Governance Practices Certification Assessment Criteria 8	**10. Procurement** Requirements: The **Participant** must have appropriate contracts in place with third-party providers of **AI Systems** and models. Contracts should, as applicable, address the following: ● a description and/or instructions on the intended use and any restrictions on use of the procured **AI System** and any data obtained in association with the procured model ● mechanisms to report on potential vulnerabilities, risks or biases that arise in the **AI System** during the tenure of the procurement agreement ● whether model training is permitted ● data protection considerations where required including any limits on the selling and sharing of **Personal Information** if applicable. Evaluation: TRUSTe must verify that the **Participant** has entered into a contract with the third party from whom the **AI System** is procured. Gaps and Remediation: If the **Participant** does not have a procurement contract, TRUSTe must inform the **Participant** that a contract with the third-party supplier of the **AI System** is required for compliance with this requirement. Where the **Participant** identifies another type of written agreement, TRUSTe must verify whether the alternative form of agreement is sufficient for compliance with this requirement.
TrustArc P&DG Control *Transparency 2.20:* Inform individuals about the ways in which data about them are processed and how to exercise their data-related rights, including those arising out of data-related incidents and breaches. EU AI Act Article 52 Transparency obligations for providers and users of certain AI systems and GPAI models G7 Code of Conduct Principle 5	**11. Privacy Disclosures** Requirement: The **Participant** must disclose to **Individuals** interacting with the **AI System** information about its use of AI inputs and outputs, and the rights of **Individuals** relating to **AI System** outcomes. Disclosures should be in clear and plain language, and inform **Individuals** of characteristics of the **Participant** in its operation of the **AI System** and of the **AI System** itself. **Individuals** should be informed that they are interacting with an **AI System** at the time of interaction, such as through a privacy notice or other mechanism (e.g. labels, disclaimers). Information can be incorporated into privacy policies that are made publicly available to all individuals who interact with the **AI System**. Evaluation: TRUSTe must verify that the **Participant**: ● informs **Individuals** that they are interacting with an **AI System** at the time of interaction; and

ISO/IEC 42001 C.2.11 Transparency and explainability NIST AI RMF MAP 2.2 OECD AI Principle 1.3 Transparency and explainability	● provides privacy disclosures about its use of AI inputs and outputs, and the rights of **Individuals** relating to **AI System** outcomes, and if permitted, that information will be used to train the AI model. Gaps and Remediation: If this information is not provided, TRUSTe must inform the **Participant** that the disclosure of information about the use of **AI Systems** and associated data is required for compliance with this requirement.

PRIVACY-ENHANCED

Implement appropriate mechanisms to limit the data used by **AI Systems** and limit the use of its outputs to the intended purposes for which **AI Systems** are to be used, and prevent re-identification of previously identifiable **Personal Information.**

TrustArc P&DG Framework and External Regulatory Standard Mapping	Assessment Criteria
TrustArc P&DG Control *Data Necessity 2.1:* Optimize data value by collecting and retaining only the data necessary for strategic goals. Leverage anonymization, de-identification, pseudonymization, and coding to mitigate data storage-related risks. GDPR Article 5.1(c) ISO 42001 A.9 Use of AI Systems NIST AI RMF MAP 1.6 OECD AI Principles 1.4. Robustness, security and safety	**12. Data Input Limitation** Requirement: The **Participant** must, where possible, limit the input of data used in the **AI System** that the **Participant** has deployed. The **Participant** shall determine and document the techniques used to ensure that the data is limited, where possible, to what is necessary and relevant to achieve the purpose for which the **AI System** has been deployed. Examples of the techniques that could be used to achieve this purpose are: ● periodic reviews of the amount of data used and nature of the inputs ● deletion of data that is no longer necessary and relevant ● limit the types of data that can be used or submitted (e.g., through settings and configurations available via the **AI System**) Evaluation: TRUSTe must verify the existence of a documented process to ensure data inputs are limited to the amount and type of data necessary and relevant to fulfill the stated purposes. Gaps and Remediation: If the **Participant** indicates it does not limit the amount data that is used as inputs, where possible, the **Participant** must be informed that these techniques must be implemented to meet this requirement.
TrustArc P&DG Control *Use, Retention, and Disposal 2.2*: Ensure data is used solely for purposes that are relevant to and compatible with the purposes for which it was collected. GDPR Article 5.1(b) ISO 42001 A.7 Data for AI Systems NIST AI RMF MAP 1.6	**13. Output Limitation** Requirement: The **Participant** must limit the use of the **AI System** outputs to the purposes for which the **AI System** is intended to be used if the output is likely to result in decisions that produce adverse legal or similarly significant effects, which may include, but not limited to, any risk to rights and freedoms of **individuals**. The **Participant** must have documented policies in place that outlines the intended uses of the **AI System**. For example: ● The output of an **AI System** that scans a candidate resume during the hiring process, must only be used for the intended purpose of helping determine the best candidate.

OECD AI Principles 1.4. Robustness, security and safety	● The output of an **AI System** that helps to determine the creditworthiness of a mortgage applicant, must only be used for the purpose of reviewing the mortgage application. Evaluation: TRUSTe must verify that the **Participant** has policies and/or processes in place to ensure that the output of the **AI System** is used for intended purpose only if the output is likely to result in decisions that produce adverse legal or similarly significant effects, which may include, but not limited to, any risk to rights and freedoms of **individuals**. Gaps and Remediation: If there are no policies and/or processes in place to ensure that the use of **AI System** outputs are limited to intended purposes only, the **Participant** must be informed that the relevant policies and/or processes must be implemented to comply with this requirement if the output is likely to result in decisions that produce adverse legal or similarly significant effects, which may include, but not limited to, any risk to rights and freedoms of **individuals**.
TrustArc P&DG Control *Use, Retention, and Disposal 2.3*: Keep data in identifiable form only as long as necessary for identified processing purposes of which individuals have been informed. If data are needed for a longer period of time for research- or optimization-related purposes, implement coding, pseudonymization, or similar mechanisms to limit the risk to individuals. NIST AI RMF MAP 1.6 OECD AI Principles 1.4. Robustness, security and safety TRUSTe Enterprise Privacy & Data Governance Practices Certification Assessment Criteria 2 and 4	**14. Prevent Re-identification** Requirement: The **Participant** must implement processes that will prevent inferences to identify Individuals or allow for re-identification of previously de-identified **Personal Information**. For example: ● the **Participant** could provide a Data Management Policy or similar policy if the document includes what measures are in place. ● a description or evidence of the process that is in place (e.g., to de-identify data). Evaluation: TRUSTe will verify whether the **Participant** has a documented process in place if **Personal Information** is used to deploy the **AI System**. Gaps and Remediation: If there are no processes in place to prevent the re-identification of previously de-identified **Personal Information**, the **Participant** must be informed that the relevant processes must be implemented for compliance with this requirement.

FAIR – WITH RISK OF HARMFUL BIAS MANAGED

Ensure accuracy of the data used by **AI Systems**, enable **Individuals** to challenge the outputs, review inputs and test for biases and accuracy, and put measures in place to review and reverse negative outputs.

TrustArc P&DG Framework and External Regulatory Standard Mapping	Assessment Criteria
TrustArc P&DG Control *Access and Individual Rights 2.13:* Enable individuals to rectify inaccurate personal data processed by the technology, process, or activity. NIST AI RMF MEASURE 2.4 OECD AI Principles 1.4. Robustness, security and safety TRUSTe Enterprise Privacy & Data Governance Practices Certification Assessment Criteria 17	**15. Data Accuracy Used to Deploy AI System** Requirement: Upon request, the **Participant** must enable **Individuals** to challenge the accuracy of the **Personal Information** that has been used by **AI Systems** deployed by the **Participant** if the output is likely to result in decisions that produce adverse legal or similarly significant effects, which may include, but not limited to, any risk to rights and freedoms of **individuals**. The **Participant** must provide access to the information used by the **AI System**, and rectify any inaccurate **Personal Information** upon receipt of sufficient information confirming the **Individual's** identity, in line with the **Participant's** individual rights request handling procedures, and applicable laws. The **Participant's** processes or mechanisms for access and rectification of inaccurate **Personal Information** used to deploy **AI Systems** must be simple and easy to use, and presented in a clear and conspicuous manner. The **Individual's** request must be responded to within a reasonable timeframe following its receipt (e.g., 45 days), and the **Individual** is provided access and rectification of their **Personal Information**, and a copy of the corrected information. If the **Participant** denies access and correction to the information used to deploy the **AI System** cannot be made, the **Participant** must explain to the **Individual** in clear and easy to understand language why the access and correction request was denied, and provide the appropriate contact information for challenging the denial of the request where appropriate. Access and correction may be denied or limited under the following circumstances: ● where providing access would violate the legitimate rights of persons other than the **Individual**; ● where the burden or expense of providing access would be disproportionate to the risks to the **Individual**'s privacy; ● where providing access would reveal the **Participant's** own confidential commercial information—such as marketing inferences,

	classifications generated by the organization, or confidential commercial information of another that is subject to a contractual obligation of confidentiality; ● where providing access would interfere with the safeguarding of important countervailing public interests—such as national security, defense, or public security; ● where **Personal Information** is being processed solely for research or statistical purposes; ● where providing access would interfere with the execution or enforcement of the law or with private causes of action—including the prevention, investigation, or detection of offenses or right to a fair trial; ● where providing access would breach a legal or other professional privilege or obligation; ● where providing access would prejudice employee security investigations or grievance proceedings or in connection with employee succession planning and corporate reorganizations; or ● where providing access would prejudice the confidentiality necessary in monitoring, inspection, or regulatory functions connected with sound management, or in future or ongoing negotiations involving the **Participant**. Evaluation: TRUSTe must verify that the **Participant** has procedures in place to respond to such requests if the output is likely to result in decisions that produce adverse legal or similarly significant effects, which may include, but not limited to, any risk to rights and freedoms of **individuals**. TRUSTe must verify that such policies, and functioning mechanisms, are available, operational, and understandable. Gaps and Remediation: If the **Participant** does not have a procedure for this, TRUSTe must inform the **Participant** that the existence of written procedures to respond to such requests is required for compliance with this requirement if the output is likely to result in decisions that produce adverse legal or similarly significant effects, which may include, but not limited to, any risk to rights and freedoms of **individuals**.
TrustArc P&DG Control *Access and Individual Rights 2.13:* Enable individuals to rectify inaccurate personal data processed by the technology, process, or activity. GDPR Article 22 NIST AI RMF MEASURE 3.3 OECD AI Principles	**16. Challenge AI System Outcomes** Requirement: Upon request, the **Participant** must enable **Individuals** to ask questions and submit complaints regarding the **AI System**, challenge its outcomes (e.g., how the decision was made and why), and request human review where the **AI System's** decisions produce adverse legal or similarly significant effects, which may include, but not limited to, any risk to the rights and freedoms of **individuals.** The **Participant** must have mechanisms and procedures in place to address an **Individuals's** questions and complaints.

1.4. Robustness, security and safety TRUSTe Enterprise Privacy & Data Governance Practices Certification Assessment Criteria 17	The request must be responded to within a reasonable timeframe following an **Individual's** complaint submission or request for human review (e.g., 45 days). Evaluation: TRUSTe must verify that the **Participant** has a mechanism to receive and procedures in place to respond to such requests. TRUSTe must verify that such mechanisms and procedures are available, operational, and understandable. Gaps and Remediation: If the **Participant** does not have a mechanism and procedure for this, TRUSTe must inform the **Participant** that the existence of a mechanism to receive complaints and written procedures to respond to such requests is required for compliance with this requirement.
TrustArc P&DG Control *Policies and Standards 1.5:* Develop policies, procedures, and guidelines to define and deploy effective and sustainable governance and controls for managing data-related risks. G7 Code of Conduct, Principle 1 ISO 42001 B.6 AI system life cycle NIST AI RMF GOVERN 1.5	**17. Inputs Review & Testing for Biases** Requirement: The **Participant** must have processes to regularly review inputs of data and test **AI Systems** to identify system biases. The quality of data used to deploy **AI Systems** potentially has significant impacts on the validity of the system's outputs. The **Participant** should implement processes to ensure that measures are integrated into the various stages (e.g., the requirement to use a specific testing tool or method to address unfairness or unwanted bias) to achieve such objectives. Evaluation: TRUSTe must verify that the **Participant** takes appropriate measures to manage data quality, including training data and data collection, to mitigate against harmful biases. Gaps and remediation: If the **Participant** does not take appropriate measures in place, TRUSTe must inform the **Participant** that the existence of appropriate measures to manage data quality, including training data and data collection, to mitigate against harmful biases is required for compliance with this requirement.
TrustArc P&DG Control *Processes 1.10:* Establish, manage, measure, and continually improve processes for establishing, implementing, publicizing, and actively managing a privacy complaint-handling process, including alternative dispute resolution as needed.	**18. Reviewing & Reversing Negative Outputs** Requirement: The **Participant** must have procedures in place to review, reverse or overturn adverse outcomes, when the outcomes result in decisions that produce adverse legal or similarly significant effects, which may include, but not limited to, any risk to rights and freedoms of **individuals.** The **Participant** must have the following in place. ● definition of what is considered a negative output ● procedures to review the complaints

ISO 42001 9.3 Management review NIST AI RMF GOVERN 1.5 OECD AI Principles 1.4. Robustness, security and safety	● procedures to evaluate and determine if a non-conformity occurred that caused negative consequences for **Individuals** ● a procedure to evaluate whether similar non-conformities exist ● a process to review the above procedures and make changes where necessary. Evaluation: TRUSTe must verify if the **Participant** has relevant processes and procedures in place to manage and reverse adverse **AI System** outputs. Gaps and Remediation: If the **Participant** does not have processes and procedures to manage and reverse adverse outputs, TRUSTe must inform the **Participant** that the existence of appropriate processes and procedures to review, reverse or overturn adverse outcomes are required to meet this requirement when decisions can produce adverse legal or similarly significant effects, which may include, but not limited to, any risk to the rights and freedoms of **Individuals.**

SAFE

Employing safety considerations prior to deploying the **AI System** can prevent failures or conditions that can render a system dangerous. Taking measures such as conducting risk assessments, pre deployment testing, and ensuring **Personnel** whose job functions that rely on the system's outputs are qualified and made aware of the implications of its outputs on **Individuals**.

TrustArc P&DG Framework and External Regulatory Standard Mapping	Assessment Criteria
TrustArc P&DG Control *Processes 1.9:* Establish, manage, measure, and continually improve processes for assessing the inherent data processing risk for new, ongoing, and modified data processing based on objective criteria for assessing risks to individuals. EU AI Act Article 29a GDPR Article 35.1 ISO 42001 6.1.2 Risk Assessment ISO 42001 6.1.4 AI System Impact Assessment NIST AI RMF MEASURE 2.10	**19. AI System Risk Assessment** Requirement: The Participant must have an AI risk assessment or similar assessments (e.g., PIA, DPIA) in place if the outcomes of an AI System result in decisions that produce adverse legal or similarly significant effects, which may include, but not limited to, any risk to the rights and freedoms of **Individuals** When the **Participant** determines that the **AI System** does not produce the above-mentioned effects, the **Participant** must provide a description of an assessment process used to make that determination. This could be included, for example, in an existing inherent risk assessment process. A risk assessment should provide the following information: ● a description of what the **AI System** will be used for ● a description of the **Participant's** processes to determine if the **AI System** will be used in line with its intended purpose ● a description of the period of time and frequency in which the **AI System** is intended to be used ● the types of data used by the system as inputs and the types of data generated by its outputs ● the categories or demographics of **Individuals** and groups likely to be affected by the system's use ● the positive impact of the **AI System** to **Individuals**, society, or the environment. ● the specific risks of harm likely to impact the identified categories of **Individuals** or groups ● a description of human oversight measures ● the measures to be taken in case that these risks materialize, including arrangements for internal governance and complaint mechanisms. The assessment should enable the **Participant** to analyze the AI risks to:

	● identify potential consequences to the **Participant** and **Individuals** affected by **AI System** outputs if identified risks occur ● determine the realistic likelihood of the identified risks ● determine the levels of risk based on the sensitivity of the data within the **AI System** and its intended uses ● determine the controls necessary to mitigate identified risks. The assessment results must be documented and made available to interested parties as determined by the **Participant** or as required by applicable law. Evaluation: TRUSTe must verify that, when applicable, the **Participant** has a process to assess, identify, document, and address the impacts and risks associated with the **AI System** outputs. When the **Participant** determines that the **AI System** does not produce above-mentioned results, a process is in place to make such determination. Gaps and Remediation: If the **Participant** does not have a process to assess and document the risks associated with the **AI System** outputs, TRUSTe must inform the **Participant** that an **AI System** risk assessment process is required for compliance with this requirement.
TrustArc P&DG Control *Resource Allocation 1.4:* Allocate appropriate resources to support the defined mission and vision, and to manage identified risks. **TrustArc P&DG Control** *Awareness and Training 1.12:* Communicate about the value and risks associated with data as well as program and process expectations. Provide both general and contextual training, including professional certification training. Reinforce messages periodically. EU AI Act Article 29 1a. ISO 42001-2023 7.2 Competence	**20. Qualified Personnel and Awareness** Requirement: The **Participant** must ensure that **Personne**l who rely on the **AI System** outputs to perform their job functions and interpret the system's outputs are competent and qualified on the basis of appropriate education, training or experience, and are made aware of the following: ● **Participant's** AI policy and/or procedures ● their responsibilities and duties relating to data and system usage, interpretation of system outputs, security, and privacy ● the implications of not conforming with the **AI System** requirements and intended use, and the impacts of decisions made based on the system's outputs. Policies, responsibilities, and implications of non-confirmatory must be communicated at the time that **Personnel** starts performing job functions that rely on the system's outputs and are reinforced on a regular basis (e.g., annually) taking into account the risks related to such tasks. Evaluation: TRUSTe must verify that the **Participant** has criteria to verify its **Personnel** are competent and qualified, and mechanisms to make end users aware of the **Participant's** AI policies, their responsibilities, and implications for not confirming with system requirements, taking into account the risks related to such tasks.

ISO 42001-2023 7.3 Awareness NIST AI RMF GOVERN 3.1 NIST 800-53 revision 5: 3.12a - Planning - PL-4. Rules of Behavior	Gaps and Remediation: If the **Participant** does not have criteria to determine **Personnel** competency and qualifications, or mechanisms to communicate policies, responsibilities and implications,TRUSTe must inform the **Participant** that competency criteria and communication mechanisms are required to meet this requirement, taking into account the risks related to such tasks.
TrustArc P&DG Control *Security 2.18:* Put in place administrative, physical, and technical safeguards to protect data from loss, misuse and unauthorized access, disclosure, alteration, or destruction. ISO 42001-2003 B.6.2.5 AI system deployment NIST AI RMF MEASURE 2.4 White House Blueprint for an AI Bill of Rights: Safe and Effective Systems	**21. Pre-deployment Testing** Requirement: The **Participant** must have processes to test the **AI System** prior to its deployment if the outcomes of the AI System result in decisions that produce adverse legal or similarly significant effects, which may include, but not limited to, any risk to rights and freedoms of **individuals**. The process should include the following: ● a testing plan outlining testing goals (e.g., release criteria) and performance metrics to be met ● both automated testing and human-led manual testing taking into account the technology being used and the role of human operators on **AI System** outcomes and effectiveness ● mirror real-world use cases in which the **AI System** will be deployed as closely as possible ● comparison of **AI System** performance against current human-driven processes. Evaluation: TRUSTe must verify that the **Participant** has a process for conducting pre-deployment testing if the outcomes of AI System result in decisions that produce adverse legal or similarly significant effects, which may include, but not limited to, any risk to rights and freedoms of **individuals**. Gaps and Remediation: If the **Participant** does not have a pre-deployment testing process, TRUSTe must inform the **Participant** that this process is required for compliance with this requirement if the outcomes of AI System result in decisions that produce adverse legal or similarly significant effects, which may include, but not limited to, any risk to rights and freedoms of **individuals**.

SECURE AND RESILIENT

Security and resilience are related but distinct characteristics. Resilience is the ability to return to normal function after an unexpected adverse event. Security includes resilience but also encompasses protecting data from loss or unauthorized use, alteration, disclosure, distribution, or access.

TrustArc P&DG Framework and External Regulatory Standard Mapping	Assessment Criteria
TrustArc P&DG Control *Security 2.18:* Put in place administrative, physical, and technical safeguards to protect data from loss, misuse and unauthorized access, disclosure, alteration, or destruction. APEC CBPR Requirement: 27 Data Privacy Framework Principles: II.4.a GDPR Article 32(1), GDPR Article 32(2) ISO 27001 8.1 Operational Planning and Control and 8.3 Information Security Risk Treatment TRUSTe Data Privacy Framework Assessment Criteria 28 UK NCSC Guidelines for Secure AI Development: Secure Deployment	**22. Security of Processing** Requirement: The **Participant** must implement reasonable physical, technical, and administrative safeguards to protect the data within the **AI System** against risks such as loss or unauthorized access, destruction, use, modification, disclosure of information, or other misuses. The **Participant** must ensure third party providers of **AI Systems** have appropriate safeguards in place if using an off-the-shelf third party **AI System**. The **Participant** must implement reasonable administrative, technical, and physical safeguards, suitable to the **Participant's** size and complexity, the nature and scope of its activities, and the sensitivity of the data within the **AI System**, in order to protect the integrity and reliability of the data and protect it from loss or unauthorized use, alteration, disclosure, distribution, or access. Such safeguards must be proportional to the probability and severity of the harm threatened, the sensitivity of the information, and the context in which it is held. These safeguards may include: ● authentication and access control (e.g., password protections, access management,limiting network and system access to authorized **Personnel**) ● **Pseudonymisation** and encryption ● implementing controls on the **AI System** query interface to detect and prevent attempts to access, modify, and exfiltrate confidential information ● boundary protection (e.g., firewalls, intrusion detection) ● physical and environmental security controls ● data backup and disaster recovery procedures ● secure data disposal procedures ● audit logging ● monitoring (e.g., external and internal audits, vulnerability scans) ● acceptable use policies that define the types of data that is prohibited to be processed through certain types of AI models (e.g., prohibition

	of using proprietary, confidential, or sensitive data through LLMs, generative AI, or similar AI technologies) - assess third party **AI Systems** to verify appropriate data protection measures are in place that are appropriate to the nature and scope of the system's activities, and the sensitivity of the data within the **AI System.** The **Participant** must periodically review and reassess its security measures to evaluate their relevance and effectiveness. Evaluation: TRUSTe must verify the existence of such safeguards and that those safeguards are adequate and proportional to the probability and severity of the harm threatened, the sensitivity of the information, and the context in which it is held. Gaps and Remediation: If the **Participant** has no physical, technical, and administrative safeguards, or inadequate safeguards to protect the data within the **AI System**, TRUSTe must inform the **Participant** that the implementation of such safeguards are required for compliance with this requirement.
TrustArc P&DG Control *Security 2.18:* Put in place administrative, physical, and technical safeguards to protect data from loss, misuse and unauthorized access, disclosure, alteration, or destruction. NIST AI RMF MANAGE 4.1	**23. Resilience** Requirement: The **Participant** must have protocols to avoid, protect against, respond to, or recover from resiliency-based threats, and procedures to regularly test these protocols post deployment of the **AI System** or confirm that existing protocols apply to **AI Systems**. Evaluation: TRUSTe must verify that the **Participant** has protocols to protect against and recover from resiliency-based threats, and procedures to regularly test these protocols after the **AI System** has been deployed. Gaps & Remediation: If the **Participant** does not have protocols and/or does not test these protocols regularly, TRUSTe must inform the **Participant** that these protocols and testing procedures are required to meet this requirement.
TrustArc P&DG Control *Security 2.18:* Put in place administrative, physical, and technical safeguards to protect data from loss, misuse and unauthorized access, disclosure, alteration, or destruction. NIST AI RMF GOVERN 1.5	**24. Incident Detection and Response** Requirement: The **Participant** must have an **AI System** incident detection, escalation, and management procedures and response plan in place, or confirm that existing incident response plans apply to **AI Systems**. Evaluation: TRUSTe must verify that the **Participant** has incident detection, escalation, and management procedures and response plan in place, and mechanisms to determine whether an incident involves **Personal Information** within an **AI System**.

NIST AI RMF GOVERN 4.3	Gaps and Remediation: If the **Participant** does not have these procedures and mechanisms in place, TRUSTe must inform the **Participant** that these procedures and mechanisms are required to meet this requirement.

III. DEFINITIONS

“AI System” is a machine-based system designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.

“Individual(s)” means the discrete person to whom the collected information pertains.

“Participant” means the entity that has entered into an agreement with TRUSTe to participate in the TRUSTe program(s) and agreed to comply with this Assurance Program Governance document and Assessment Criteria of the program(s) in which the **Participant** is participating, and is the party who is either utilizing a third-party **AI System** and/or deploying its own built/home-grown/open-source solution.

“Personal Information” means any information about an identified or identifiable **Individual**, including indirect identification of an **Individual** through an identifier (e.g., identification number, location data, or online identifier) or through other factors (e.g., genetic, physical, or social identity).

“Personnel” is a group of **Individuals** that work for a company or institution.

“Pseudonymisation” is the processing of **Personal Information** in such a manner that the **Personal Information** can no longer be attributed to a specific **Individual** without the use of additional information, provided that such additional information is kept separately and is subject to technical and organizational measures to ensure that the **Personal Information** is not attributed to an identified or identifiable natural person.