Responsible Artificial Intelligence Checklist

This checklist aims to help developers and deployers of artificial intelligence (AI) systems, including deployers of AI systems procured from a third-party, to examine their AI-related activities to determine if they meet applicable best practices.

Note: Deployers of third-party AI systems are still accountable for the AI systems they deploy. The best practices outlined in this document still apply to third-party AI systems that your organization procures and deploys and should be considered to the extent that your organization has the capability to comply with them.

Table of Contents

TrustArc

AI Governance

- [ ] Incorporate responsible AI practices into the organization's digital responsibility strategy with a focus on taking responsibility for artificial intelligence (AI) systems developed or used by the business

- [ ] Determine who has ultimate responsibility and accountability for the business' AI decisions

- [ ] Define and implement internal policies and guidelines around how the organization will use and govern AI systems:
    - Draw from the OECD AI Principles[1] (and other authoritative AI principles and standards):
        - Inclusive Growth, sustainable development, and well-being
        - Human-centered values and fairness
        - Transparency and explainability
        - Robustness, security, and safety

- [ ] Develop and apply standardized processes and rules to develop, test, deploy and operate AI systems (e.g., coding, documentation, testing, architectural guidelines):
    - Review guidance from relevant regulators, advocates, and industry leaders

- [ ] Conduct internal audits to assess adherence of the organization's AI practices to the standardized processes

- [ ] Designate personnel responsible for keeping the business current on regulatory and technical developments:
    - Revisit internal policies and procedures to ensure that they address and account for any changes in the AI regulatory landscape, including:
        - Privacy policies
        - Data use policies
        - Information classification and management policies
        - Terms of service
    - Ensure employees know who to contact with questions, inquiries and/or concerns about the business's privacy and data protection policies regarding AI.

[1] OECD AI Principles and Recommendations for Policymakers https://oecd.ai/en/ai-principles

AI Leadership

- [ ] Designate a member of senior leadership/management to be responsible and accountable for governance of the business' development and use of AI systems, including enforcement of internal policies and procedures

- [ ] Make the AI leader accountable for working with the business' Data Ethics Committee (or equivalent)

- [ ] Clearly define all organizational roles and responsibilities associated with the design, development, use and deployment of AI systems to ensure such roles and responsibilities align under the AI leader's portfolio

Data Ethics

- [ ] Distinguish between the business' ethical and legal responsibilities:
    - A practice may make the business legally compliant with a regulatory obligation (e.g., obtaining consent), but this does not make such practice, or the way it is carried out, automatically ethical

- [ ] Develop a set of AI ethical principles and an internal Code of Conduct/Business Code of

- [ ] Conduct that makes developers and users of AI systems accountable for compliance with the AI ethical principles:
    - Draw from the OECD AI Principles[2] (and other authoritative principles)
    - Integrate the AI principles into the Code of Conduct/Business Code of Conduct and vice-versa
    - With transparency in mind, develop internal ethical guidelines that align to the business' AI ethical principles in a manner that they can easily be made public should the need arise

- [ ] Establish a Data Ethics Committee (or equivalent) to advise the business on the ethical development and use of AI systems:
    - The Committee should act in an advisory capacity only, but should not be easily overruled by operations or leadership

- [ ] Comprise the Data Ethics Committee of individuals and roles from all levels of the company:
    - The Committee should be comprised of senior leadership, C-suite or board members

- [ ] Include on the Data Ethics Committee individuals with diverse backgrounds and contrasting opinions

[2] OECD AI Principles and Recommendations for Policymakers https://oecd.ai/en/ai-principles

- [ ] Provide a “safe space” for Committee members to challenge business practices and make recommendations to the business

- [ ] Make issues and topics discussed accessible to all Committee members (e.g., those that do not have a legal or operational background)

- [ ] Distinguish between ethical and legal responsibility:
    - A practice may make the business legally compliant with a regulatory obligation (e.g., obtaining consent), but this does not make such practice, or the way it is carried out, automatically ethical

Ethics by Design

- [ ] Identify which key stakeholders should be involved in designing, developing and implementing responsible AI solutions

- [ ] Establish what “fairness” means to the stakeholder group. This can be done with guidance from the Ethics Committee

- [ ] Clearly define and articulate the business value or context of business use for the use of an AI system:
    - Start with a problem statement (i.e., a problem that the business is trying to solve):
        - Avoid starting with a solution or capability that the business has, otherwise this may force an AI solution where one may not be warranted
    - Document underlying assumptions about how the AI system will operate so that these assumptions can be validated through monitoring and testing

- [ ] Define the terms, processes and data sets needed to build the solution:
    - If Privacy Enhancing Technologies (PETs) are used, integrate them all the way through the development lifecycle rather than trying to fit them in somewhere at the end.

- [ ] Pilot the AI system before deploying it in production to ensure it is operating as intended

- [ ] Determine whether the business is engaging in good practices and what are the consequences of certain practices

- [ ] Engage with the business’ Ethics Committee to ensure stakeholder groups and the Ethics Committee use consistent language and terms, and are not working against each other to develop and produce ethical algorithms

- [ ] Put in place organizational policies and practices to foster critical thinking and a safety-first mindset in the design

- [ ] Consider risks associated with inputting sensitive information into AI models:
    - Generative AI models should not be fed or prompted to output sensitive or confidential information unless data is processed locally and/or subject to appropriate access and use controls
    - Special care should be given to the following:
        - Children's data
        - Employee data
        - Education data
        - Healthcare patient data
        - Hiring or workplace data that could lead to claims of discrimination or harassment
        - Other regulated forms of data

- [ ] Establish mechanisms and processes to demonstrate that the AI system being deployed is valid and reliable, e.g., with regard to:
    - Data collection
    - Lawfulness of processing
    - Limiting processing to the specified purpose

- [ ] Establish mechanisms for personnel designing or deploying AI systems to regularly incorporate feedback from the Ethics Committee and relevant AI actors (e.g., software developers, end users, operators, data scientists, legal and compliance, etc.) into the system's design and implementation

- [ ] Put in place contingency processes to handle failures or incidents in third-party data or AI systems deemed to be high-risk:
    - Ensure the AI system can fail safely, such as when made to operate beyond its knowledge limits, and that residual negative risks do not exceed the business's risk tolerance

Data Governance

- [ ] Implement data governance policies and procedures that set the standard for how the organization will collect, use, store, maintain, and disseminate personal data related to the use of AI systems[3]

- [ ] Document and keep up to date data provenance records that address the following:
    - The categories of data collected (personal and non-personal)
    - The origin/source of the data
    - Original purpose of collection
    - Clear explanations of what data is used, how it is collected and why
    - Who labeled the data and whether bias tests were conducted to assess if the labeled data was biased
    - How data is transformed over time
    - Date of last update or modification (this can include the date tag in metadata)

[3] See Sample Data Governance Policy under Privacy Management Category 1 - Integrate privacy into Data Ethics/Stewardship program

- [ ] Implement mechanisms to enable individuals to challenge the accuracy of personal data used by AI systems deployed by the organization:
    - Rectify any inaccurate personal data upon receipt of sufficient information confirming the individual's identity, in line with the organization's procedures for handling privacy-related requests and all applicable laws
    - Make mechanisms simple and easy to use, and present them in a clear and conspicuous manner

Data Input and Output

- [ ] Limit the input of personal data used in the AI system:
    - Review and update existing data minimization policies and procedures to address limiting the input of data used in AI systems
    - Determine and document the techniques used to ensure that data used as AI system inputs is limited, where possible, to what is necessary and relevant to achieve the purpose for which the AI system has been deployed. Techniques may include:
        - Periodic reviews of the amount of data used and the nature of the inputs
        - Deletion of data that is no longer necessary and relevant
        - Limiting the types of data that can be used or submitted (e.g., through settings and configurations available via the AI system)

- [ ] Limit the use of 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, including any risk to the rights and freedoms of individuals
    - Review and update existing policies addressing purpose limitation to address limiting the use of AI system outputs
    - Develop and document policies that outline the intended uses of the AI system

Procurement

Third-party AI system procured for deployment by your organization:

- [ ] Develop criteria (in collaboration with procurement and enterprise risk management teams) to assess and approve new or updated third-party software and services that integrate with AI APIs or offer AI features:
    - Have internal reviewers consider the data sets used to create the outputs (e.g., for generative AI models) to determine the unique risks associated with each software and service
    - Review outputs for explainability, fairness and bias – do the outputs make sense based on the inputs, are the outputs fair and nondiscriminatory
    - Assess the terms of use and system documentation of the third-party vendor; are they sufficiently transparent about the AI features

- [ ] Undertake to understand how the third-party AI system is built, including:
    - The data used to train the AI system
    - How the AI system is assessed by the developer for effectiveness and explainability
    - Under what circumstances does the AI system perform poorly

- [ ] Ensure contracts with AI system third-party providers that address the following:
    - A description and/or instructions on the intended AI system
    - Any data obtained in association with the procured model
    - Mechanisms for each party to report to the other on potential vulnerabilities, risks or biases that arise in the AI system during the tenure of the procurement agreement

Engaging a third-party service provider that uses an AI system to deliver services:

- [ ] Establish requirements for sharing data with vendors that ensure compliance with relevant laws, regulations and best practices for sharing or the sale of data to third parties:
    - Minimize the sharing of personal data as part of data sets used to train third party AI systems

- [ ] Review contractual terms to ensure that any uses of data by vendors reflect mandatory regulatory contractual language, or are subject to approved exceptions:
    - As required by local law, ensure the appropriate notices and consent are obtained prior to sharing personal data with vendors
    - Ensure vendors are able to comply with individual requests regarding the use of their personal data in systems using AI (e.g., Do Not Sell requests under US state laws, access and deletion requests)

Transparency and Explainability

- [ ] Examine and document risks associated with transparency and explainability:
    - Determine how the business should provide transparency to the public or impacted individuals about its use of AI
    - Determine when and how employees are required to disclose whether internal and/or external work product was created in whole or in part by AI tools, particularly generative AI tools

- [ ] Inform individuals when and how they are interacting with AI:
    - Inform individuals that they are interacting with an AI system at the time of interaction, such as through a privacy notice or other mechanisms (e.g., labels, disclaimers)
    - Provide information in a clear and concise manner
    - Make the information easily accessible
    - Clearly indicate to individuals when they are interacting with a machine (bot) or automatically generated content

- [ ] Inform individuals that interact with the AI system of their rights relating to the AI system's outputs and outcomes

- [ ] Document and make accessible to impacted individuals/communities and the public, where appropriate, explanations of how the AI model works, including:
    - Mechanisms underlying the AI system's operation
    - The AI system's knowledge limits

- The data the system collects and/or uses
- Reasoning for all outputs generated by the system
- How system outputs may be usedAny decisions being made and the parameters for how such a decision is to be reached
- How the system produces different outcomes
- The involvement of human interveners (e.g., how/where in the system's processing cycle does human intervention occur and to what degree)

☐ Make explanations appropriate to the stage of the AI lifecycle and risk level of the AI system, and tailor explanations 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
- Feature importance

☐ Ensure explanations provided about the AI system match the processes used by the system for generating the output:
- Make explanations meaningful:
  - Use language, concepts and terms most understandable to the particular audience of the explanation (i.e., the same explanation may be provided with different levels of complexity depending on the audience)
- Confirm that the system is only used for what it was intended

☐ Consider making the algorithm's source code publicly available, or at the very least, available upon request

☐ Establish processes for end users and individuals/communities impacted by AI decisions to report problems and appeal system outcomes;

☐ Communicate incidents and errors to relevant AI actors and impacted individuals/communities

☐ Be transparent about the design of AI systems – disclose information about:
- Design goals
- Data inputs
- The construction and operation of the system
- System outputs and their impacts on targeted individuals/communities and/or society as a whole

Interpretability

- [ ] Determine what information needs to be interpretable so that the design and outcomes of the system can be easily explained:
    - The information needed for interpretability will depend on the stakeholders and context of accountability demands

- [ ] Regularly review documentation explaining how the AI system works and makes decisions/generates outputs, and update the documentation where necessary. Documentation should address the following:
    - The time tracking of reviews
    - Personnel involved in the review
    - What decisions/outputs are reviewed
    - What updates are made, if necessary

- [ ] Make AI systems auditable:
    - Enable third parties to probe, understand and review AI systems before deployment and during operation.

Impact Assessment

- [ ] Implement processes to assess, identify, document, and address the impacts and risks associated with automated decision-making and AI system outputs:
    - Include the following in the assessment:
        - A description of what the AI system will be used for
        - A description of the organization'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.

- [ ] To effectively analyze risks to the organization, the assessment should:
    - Identify potential consequences to the organization and individuals affected by AI system outputs should an identified risk occur
    - 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

- Make the assessment available to interested parties and relevant stakeholders as determined by the organization or applicable law

☐ Prioritize identified AI risks based on impact, likelihood and available resources or methods

☐ Develop, plan and document responses to AI risks deemed high priority:
- Risk response options include mitigating, transferring, avoiding or accepting the risk

Human Oversight and Intervention

☐ Document the AI system's knowledge limits and how system outputs may be used and overseen by humans (i.e., human intervention)

☐ Implement mechanisms to alert human operators to adverse outcomes or impacts of the AI system

☐ Implement mechanisms for human oversight and intervention where the purpose of processing is likely to 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, particularly where the AI system:
- Cannot detect or correct its own errors
- Requires human oversight for the AI system to function properly according to its instructions or other documentation for its intended use

☐ Define, assess, and document mechanisms for human oversight and intervention in accordance with organizational policies

☐ Determine the right level of human oversight and intervention needed:
- Human operators must have the authority to override errors and decisions made by the AI system
- Human interveners must have clear and documented processes and be supported by interpretability mechanisms

☐ Document detailed processes for human interveners on making decisions and taking required subsequent actions, including:
- How a human operator is notified of/becomes aware of situations requiring intervention (e.g., proactive checks, regularly monitors outputs for inaccuracies, random audits)
- Procedures for how a human operator corrects errors in the AI system

☐ Implement mechanisms and capabilities for individuals impacted by the AI system's outputs to report adverse impacts (e.g., unfairness) and/or request human intervention

Training and Awareness

- [ ] Communicate to and train employees on their responsibilities when working with AI systems at the time that they start performing job functions that rely on the AI system's outputs, including:
    - The implications and consequences of using AI tools in the workplace (e.g., outputs of certain tools can be inaccurate):
        - Keep in mind that the use of AI and concept of ethics may be a novel concept to some employees
    - 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's requirements and intended use, and the impacts of decisions made based on the system's outputs

- [ ] Clearly communicate to employees if there are tools specifically recommended or prohibited:
    - Provide clear guidance on how employees can or cannot use AI tools to perform their essential job functions
    - Make employees aware of relevant policies, such as the organization's AI policy or Acceptable Use Policy

- [ ] Remind employees that relevant legal obligations continue to apply to the use of new tools;

- [ ] Warn employees against inputting personal data, confidential business information, trade secrets or other sensitive data into AI systems

- [ ] Consider specific training to help employees understand and mitigate legal liability in the use of certain AI tools, particularly for businesses in a regulated industry
    - Update employee resources, including employee handbooks to reflect policies regarding AI use

- [ ] Establish and enforce development guidelines to hold employees dealing with AI systems accountable

- [ ] Establish and communicate protocols for employees using AI applications on work-issued devices:
    - Advise employees which settings or permissions are acceptable

- [ ] Provide employees with resources that address the responsible use of all types of AI

- [ ] Hold training sessions and workshops on different aspects of AI, such as:
    - Ethics and bias
    - Data minimization
    - Data accuracy/inaccuracy
    - Transparency and explainability
    - Data security

- [ ] Provide specialized training for employees responsible for human intervention processes: Develop, document and train all human intervention processes before the launch or use of an AI system so that customers can exercise such an alternative from day one

- [ ] Reinforce training by setting up refresher training at regular intervals to keep employees up-to-date on current legal restrictions/permissions and emerging risks associated with AI systems (e.g., annually)

Monitoring and Control

- [ ] Identify and assess how the business will measure AI risks (e.g., most significant to least significant) and risks that cannot be measured must be properly documented:
    - Implement safety metrics that reflect:
        - System reliability and robustness
        - Real-time monitoring
        - Response time for system failures
        - Connect measurement approaches for identifying AI risks to deployment context(s)

- [ ] Implement post-deployment plans to continually monitor the functionality and behavior of AI systems and their components to ensure they are performing as intended during the tenure of deployment:
    - For third party systems, determine whether the AI system is performing according to the intended us as described by the AI developer
    - Confirm that outputs are consistent with the organization's expectations
    - Integrate feedback and problems reported by end users and individuals/communities impacted by AI decisions into evaluation metrics
    - Implement policies and procedures that address the following:
        - Who will be involved in the monitoring and analysis
        - What needs to be monitored
        - The methods needed for monitoring, analysis and evaluation
        - The frequency of the monitoring
        - The expected performance according to intended use
        - When monitoring and analysis will be performed
        - Defined metrics relevant for monitoring

- [ ] Determine the course of action to be taken if AI systems do not perform according to their intended use, e.g.:
    - Update and retrain the AI system's inputs and/or outputs so that the system performs according to its intended use (to the extent the organization has control over the inputs for third party systems)
    - Cease the use of the AI system where outputs are likely to result in decisions that produce adverse legal or similar effects of similar importance, and in the case of a third-party system, notify the developer that the system does not perform to its intended use

- ☐ Regularly assess the metrics used to measure AI risks for effectiveness and update them as necessary:
    - Assessments should be conducted by independent assessors or internal experts who were not involved in the system's design and development)

- ☐ Implement mechanisms and capabilities for external parties to report adverse impacts (e.g., unfairness):
    - Have procedures in place to review, reverse or overturn adverse outcomes, especially those that result in decisions that produce adverse legal or similar effects of similar significance, including any risk to rights and freedoms of individuals – such as:
        - Procedures to review complaints
        - 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

- ☐ Obtain and document feedback from domain experts and relevant AI actors on measurement results regarding AI system trustworthiness in deployment context(s) and across the AI lifecycle to validate whether the system is performing as intended

- ☐ Implement AI system incident detection, escalation, and management procedures and response plan, or update existing procedures and response plan to apply to AI system

Testing

Pre-deployment:

- ☐ Implement processes and procedures to test the AI system prior to its deployment. Policies and procedures 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
    - Real-world use cases in which the AI system will be deployed – mirror these use cases as closely as possible
    - A comparison of AI system performance against current human-driven processes

Post-deployment:[4]

- ☐ Regularly review inputs of data and test the AI system to identify system biases:
  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)

- ☐ Implement mechanisms to audit the accuracy of the AI system's outputs in relation to the intended performance of the system:
    - E.g., implement testing procedures to measure the accuracy of the AI system's outputs that address the following:

[4] See AI Testing Procedures under Privacy Management Category 4 - Maintain policies/procedures for algorithmic accountability

- A description of the methodology used to measure and monitor the accuracy of AI outputs
- Clearly defined and realistic test sets that are representative of the conditions of intended use
- False positive and false negative rates
- Human involvement, if any, in AI decision-making

- Testing mechanisms should be in place to assess the accuracy and quality of generated outputs when new inputs are added

☐ Develop methodologies and technical measures to test AI systems to detect and address sources for negative outcomes:

- Validate assumptions made about the AI system in the design phase, or if using a third-party AI system, in the pre-deployment testing phase
- Confirm that data collection and/or processing by the system is consistent with the intended purpose of the system
- Validate that the impact of outcomes/decisions made by the AI system are not harmful to any individual, community or society as a whole

☐ Recalibrate the system based on results of the testing to ensure ongoing compliance with ethical and legal requirements

Incident Detection and Response

☐ Review and update existing incident detection, escalation, and management procedures, including the organization's incident response plan, to incorporate AI-related incidents:

- Implement mechanisms to determine whether an incident involves personal data within an AI system

Resources

Algorithmic Accountability – David Horneber and Sven Laumer, Springer Link
https://link.springer.com/article/10.1007/s12599-023-00817-8

Generative AI for Organizational Use: Internal Policy Checklist – Amber Ezzell, Future of Privacy Forum, July 2023
https://fpf.org/wp-content/uploads/2023/07/Generative-AI-Checklist.pdf

Lobschat L, Mueller B, Eggers F, Brandimarte L, Diefenbach S, Kroschke M, Wirtz J (2021) Corporate digital responsibility. J Bus Res 122:875–888.
https://doi.org/10.1016/j.jbusres.2019.10.006

Schneider J, Abraham R, Meske C, Vom Brocke J (2022) Artificial intelligence governance for businesses. Inf Syst Manag. https://doi.org/10.1080/10580530.2022.2085825

Donia J (2022) Normative logics of algorithmic accountability. In: ACM conference on fairness, accountability, and transparency, pp 598–598.
https://doi.org/10.1145/3531146.3533123

Ensuring individuals can fully exercise their rights - 21 September 2022 - CNIL.
https://www.cnil.fr/en/ensuring-individuals-can-fully-exercise-their-rights

NIST AI Risk Management Framework 1.0 - NIST AI 100-1 -January 2023.
https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

TRUSTe Responsible AI Certification Assessment Criteria