Evolving to An Effective Algorithmic Impact Assessment

Peter Cullen
Executive Strategist
Information Accountability Foundation[1]

Abstract

As more companies rely on Artificial Intelligence (AI) to make critical decisions, the risk from the use of AI increases. Academics, NGOs and some policymakers increasingly are recommending using Algorithmic Impact Assessments (AIAs) as part of enhanced governance systems to assess the potential benefits, risks, and controls with the goals of achieve responsible and ethical AI. An AIA is another name for a more expansive impact assessment outlined in the IAF's report AI and the Road to Expansive Impact Assessments. They are broader than, for example, what is required in Article 35 of the GDPR relating to Data Protection Impact Assessments (DPIA's) and what is outlined in the EU's proposed AI Regulation relating to conformity assessments.

To date, there is no consensus on the structure and focus of an AIA, although many NGO organizations have released high-level proposals for structure. These high-level proposals, though valuable to the discussion around AIAs, have not addressed all key elements for a successful assessment. To mirror existing assessments, a fully functional AIA should have several supporting and necessary parts. A set of interrogative and iterative questions need to be answered as part of the development process for any AI solution (AIS). It should start with an assessment of the initial level of risk that would determine what level of review and oversight would be required at each stage of the AIS development lifecycle. As the development process for AI is more complex and iterative than traditional software, a defined program management system and enhanced governance processes also are required. Key stage or review gates need to be established at specific points in the development cycle.

To address these areas, an approach to an AIA is outlined, as part of a governance system that includes a risk tiering approach that helps a higher risk AIS receive a more detailed review at key stages. A model set of interrogative questions that build to a decision point and key questions to ask as part of the defined stages of the review process are provided. The shortcomings of existing assessment processes, such as DPIAs, suggest an AIA can be built by expanding existing assessment processes, including a DPIA. Finally, changes or enhancements to existing governance systems that provide mechanisms to meet the external ethical use expectations of AIS are addressed.

Organizations need to adopt and prepare for additional governance requirements that include more expansive AIAs to satisfy demonstrable accountability expectations that increasingly are demanded by policy makers, regulators, consumers, and other stakeholders[2].

1 Background and Motivation

In both commercial and public sectors, AI use is expanding rapidly within organizations. Common uses of AI are automating existing processes, augmenting decision-making, creating new products and services, and enhancing customer experiences. To effectively monitor these expanded uses of AI, organizations quickly need to explore ways of effectively overseeing AI application risk from development through deployment. AI systems are complex, making then harder for organizations to understand both expected and unintended impacts. Unintended consequences can yield risks and potential harms for both consumers and organizations using and deploying AI.

The complexity and risk of the AI systems increase the type and breadth of the governance[3] processes needed. The explanations associated with these systems and the increasing complexity of the AI development chain (e.g., the data, data use and technology of AI often involves a wide combination of partners, vendors, and providers) add to the necessary governance.

---

[1] See appendix lead for acknowledgements to contributors.
[2] Ten Principles of Responsible AI for Corporates | by AnandSRao | Towards Data Science
[3] The term "governance" refers to the collective set of policies, procedures, and oversight internally and externally that manages the risk of systems and meets required obligations,

The clear development stage of AI, where models are built and tested, is distinct from the deployment stage. This distinction introduces iterative analysis requirements and decisions and discrete program management requirements.

Some of the new or expanded set of risks of AI applications (Figure 1) that organizations need to address are localized to individual applications, like those of bias and discrimination or lack of interpretability. Other risks are much more societal and global, reflecting the potential for systemic discrimination, exacerbation of inequalities and job displacement, and broader ethical risks, like surveillance. The breadth of these risks highlights the need for additional governance, including risk identification, oversight, and controls within organizations adopting these complex systems.

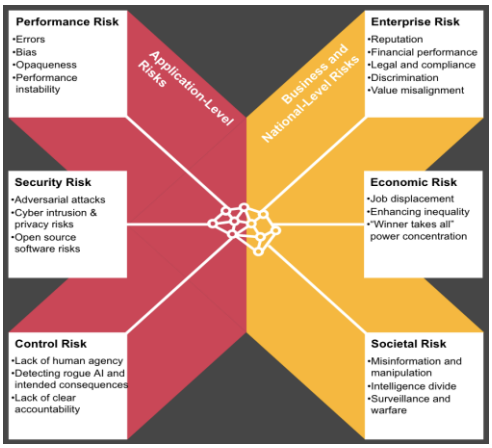


Figure 1: PwC Taxonomy of AI Risks

1.1 Legislative and Regulatory Landscape

Legislation and regulations have not kept pace with these challenges and risks. While some existing laws remain relevant, like the Equal Credit Opportunity Act in the US, there remains a gap in the laws and regulations intended to meet the challenges of the AI environment. This is changing but in a sporadic manner.

There are AI-specific proposed bills in the U.S. and globally, and there has been much governmental debate and investigation as to how to govern the impact (risks) of AI. These bills universally contemplate some form of impact assessment and increasingly algorithmic audits. Several proposed laws in the U.S. introduce specific requirements related to algorithmic assessments or risk assessments related to "automated decisions" which is becoming the proxy integration medium between AI and privacy. The growth in the use of AI likely will accelerate this trend toward impact assessments in general.

Proposed specific assessment laws in the U.S. Senate include the Algorithmic Accountability Act and the Algorithmic Fairness Act. Also in the U.S. Senate are proposed privacy bills with either the explicit or implicit requirement for much broader type risk assessments (e.g., the Mind Your Own Business Act and the Consumer Online Privacy Rights Act). In U.S. state legislatures are the Automated Decision Systems Accountability Act of 2020 (California), as well as the Virginia and Colorado privacy laws which have similar assessment requirements.

The EU's proposed Regulatory Approach to AI contains specific risk assessment requirements, but it requires the performance of "conformity assessments."[4] These generally map to established regulations and standards relating to product safety and which by themselves do not address the full range of risks raised by AI.[5] DPIAs are required by the EU General Data Protection Regulation (GDPR), but DPIAs insufficiently address the complexity and risks AI solutions pose.

2 Why DPIAs and the Existing Proposals are not Enough

PIAs must be conducted by many governmental agencies (i.e., U.S. E-Government Act of 2002; Canadian Directive on PIAs. This public sector requirement also has been imposed on the private sector in some countries. As mentioned above, a DPIA is required under the GDPR (Article 35) when the data processing is considered "risky". Already, most AI applications will require a DPIA under the GDPR because they will trigger a "risky processing" determination in at least two of the nine DPIA requirements as outlined in guidance by the European Data Protection Board (i.e., "new or innovative use of technology, data processed on a large scale, automated decision making, evaluation or scoring"). This perspective recently was reinforced by the UK ICO.

[4] Conformity assessment | NIST

[5] See the IAF's response to the EU Proposed AI Regulation for further details on Conformity Assessments vs AIA

DPIAs are limited because they focus on the individual as opposed to a broader set of stakeholders, and then, only if the technology involves the use of personal data. Even if, for example, personal data were used in an AI application, such as a simple automation that could result in a job role change or even a job loss, a DPIA may not address this potential impact.

AI stretches the application of personal data. In some settings, AI may make non-personal data identifiable in two ways. First, it broadens the type of and demand for collected data (e.g., sensors in cell phones, cars and other devices). Second, it infers complex relationships from that data (e.g., facial features, gait, fingerprints, and other forms of biometric recognition technologies). This expanded set of discrete data can be combined in a way to infer highly sensitive and accurate information about an individual. For example, researchers were able to take a person's Facebook likes and infer the person's gender, sexuality, age, race, and political affiliation based solely on these surface data.[6] These data and uses should be part of an impact assessment but currently may not be explicitly covered by a DPIA.

The recitals and guidance on DPIAs suggest the rights considered should include economic situation, health, personal preferences or interest, exclusion, or discrimination in addition to more standard data protection rights. Whether an impact assessment associates or covers all fundamental human rights is unclear. Recital 4 of the GDPR comes closest to addressing a broad range of rights and interests; however, it was never carried through to the full text of the GDPR. In addition, the evaluation of risks required in an assessment such as a DPIA does not address the risk of **not receiving** a "benefit" (e.g., increased access to health). In short, a DPIA falls short in enabling an organization to balance interests across all stakeholders and evaluate the ethical impact of AI and associated data use.

As a DPIA is nested inside the GDPR, there are several terms that are not fully suitable in an AI environment. For example, "transparency" and "accuracy" have defined meanings in the GDPR but do not align with the way the terms are used in AI or do not adequately address the associated impact. As noted in the IAF blog Transparency Needs a Makeover, the current focus on the right to explanation is too narrow. Transparency, in a GDPR context, is narrower than what is thought about in an AI context. Accuracy in data protection is one of the fundamental principles, focusing on the accuracy, quality, and maintenance of personal data. Broadly, accuracy in AI refers to how often an AI system produces the correct outcome, measured against correctly labelled test data. These differing definitions may cause confusion between privacy and data science teams.

But a core reason a DPIA is not sufficient in an AI context is that AI can bring added risks. By design, DPIAs are intended to, or have been implemented to, primarily address compliance with privacy regulations. By extension, they largely do not address other AI-specific risks and impacts and the associated governance requirements of AI. However, they can serve as a building block to a more expansive AI assessment.

A new assessment type increasingly is gaining momentum, the AI or Algorithmic Impact Assessment (AIA).

Few proposed approaches for AIAs consider the relationship or potential overlap with DPIAs. Also, there does not seem to be a consensus on what an AIA should include and address. The AIA and DPIA are both risk assessment tools and partly use the same logic. Both instruments are complementary in terms of meeting existing and emerging regulatory requirements but are not interchangeable. To date, only DPIAs and PIAs are mandated, and then only under certain contexts.

Few suggested AIA structures attempt to build off or improve the existing DPIA mechanism, and most that do so are in their infancy.[7] Many do address a broader range of issues than a DPIA (e.g., influencing human behavior e.g., addictive or the impact to the environment) and address a broader range of stakeholders (e.g., society, democracy). Some focus too much on the technology and its impact and too little on actual means to assess the impact of data use. Few address the practical realities of data governance/management required for the effective governance of AI. Most do not propose a way to identify impacts

---

[6] Rosen, R.J. 2013. Armed With Facebook "Likes" Alone, Researchers Can Tell Your Race, Gender, and Sexual Orientation. *The Atlantic*.

[7] The algorithm audit: Scoring the algorithms that score us - Shea Brown, Jovana Davidovic, Ali Hasan, 2021 (sagepub.com) | AI System Ethics Self-Assessment Tool: 2019. *https://www.smartdubai.ae/self-assessment.*| | Recht, A.I.& 2019. *Artificial Intelligence Impact Assessment -Netherlands* | Reisman, D. et al. 2018. Algorithmic impact assessments: A practical framework for public agency accountability. *AI Now Institute*. April (2018), 22.

and risks in a programmatic way and do not offer a way to translate ethical values into measurable risks. They also tend to include inadequately a balancing function and an opportunity to outline mitigating controls for identified risks. Importantly, given they do not leverage the DPIA itself, these AIAs introduce yet another compliance mechanism and burden to organizations. Proposed primarily by the data science and academic communities, proposed AIAs contain some of the same concepts though are intended more as a mechanism for documentation and review rather than compliance or conformance to ethical commitments.

An AIA can complement other proposed mechanisms and aggregate outputs from risk tiering, checklists, metadata documentation, and ethical AI software tools engaged throughout the AI lifecycle to perform a balancing interests exercise. They also could incorporate conformity assessments as set forth in the EU Proposed AI Regulation.

3.0 A new Proposed AIA Structure

An AIA should cover the full development lifecycle requirements from strategy and planning, model development, the specific issues related to training data, deployment, ongoing operation and monitoring issues, and governance. It should encompass consideration of not just the technology itself but also core data and data use governance that reflects the interests of all stakeholders. It should address the impact of the use of data, irrespective of whether the data used in development were personal data or not. It should address facets of model risk management.[8] It would provide for a demonstrable decision-making framework to enable not only responsible AI but also to serve as an explanatory mechanism to internal and external stakeholders through documentation of the review and approval of an AI application. It should include evidence for audit purposes, grounds for design requirements and governance actions, and feedback loops. Not only would such an AIA approach meet the requirements of ethical and responsible AI, but it could also subsume the balancing and assessment requirements of a DPIA.

An AIA needs to expose both the risks and the benefits and facilitate a trade-off analysis between the two. The AIA also should include an overview of the controls engaged and should outline the final decision made by an organization to move or not move forward with an application.

Business objectives, culture, risk profiles and each AI application for a given business can be unique. This uniqueness means an AIA cannot be standard and requires modification to suit the organization's environment. At a high-level, the design of a specific AIA process can be broken down into several parts.

First, is the “risk gating criteria” which can serve at least two purposes:

1. Not every AI application will require the same scrutiny (e.g., a simple workflow automation project will likely create fewer potential risks than the use of facial recognition data in an AI System (AIS)). A set of risk criteria can ensure the right level of scrutiny is allocated to higher potential risk applications.
2. Related to levels of risk, a stacked criteria list also can enable decision-making; higher risk applications should be reviewed by more senior organization approvers. While there is no consensus as to what AIS may create higher risk and it may depend on the specific business or use, as an example, the following criteria could be used to flag an AIS as higher risk:

    - Involves financial or other regulatory compliance decisions or direct human or decision impact to healthcare or HR
    - Use of large-scale personal data and/or human characteristics data or decision (observable human characteristics for example keyboard logging) and/or sensitive data
    - Usage of facial recognition or any biometric data
    - Socio economic, political, and reputational integrity, environmental or physical impact (e.g., autonomous driving)
    - High degree of AI autonomy that can have a significant impact on persons or entities or that have legal consequences for them
    - AI is applied in a new (social) domain

[8] What is Model Risk Management and How is it Supported by Enterprise MLOps (dominodatalab.com)

Second is the set of questions an assessment should cover. The initial question set of an AIA should address the high-level purpose of the application and the accountability for the project (e.g., "What are the objectives of this initiative for each group of stakeholders" and "Who has ultimate decision-making authority").

Third are the questions related to the planning, development, and deployment of the AI application. Many of these questions are an assessment of risk since the way a particular requirement, such as data testing, is adopted can either increase or mitigate risk. These questions should match to standards or sets of operating procedures established as part of the governance process. Questions addressing traceability of data, model versioning and usage, capturing performance against success criteria, types and sources of data that will be used, testing processes to evaluate fairness and bias, and specific questions regarding whether third-party data or technology should be used.

Fourth, arguably the most important as it assesses "impact," are the questions that explore fully a multi-stakeholder analysis of all benefits, risks and mitigating controls. They should address the likelihood and significance of benefits and risks and the effectiveness of controls to reduce risk. This analysis should be guided by the way requirements were addressed as part of the likelihood of both benefits and risks materializing and the application of associated controls.

Finally, are questions addressing the go/no go decision by the accountable parties and any additional factors or controls that are required as part of the approval process.

Appendix 1 outlines a model AIA questions set. It includes many of the element areas outlined in the Machine Intelligence Garage Ethics Framework.[9]

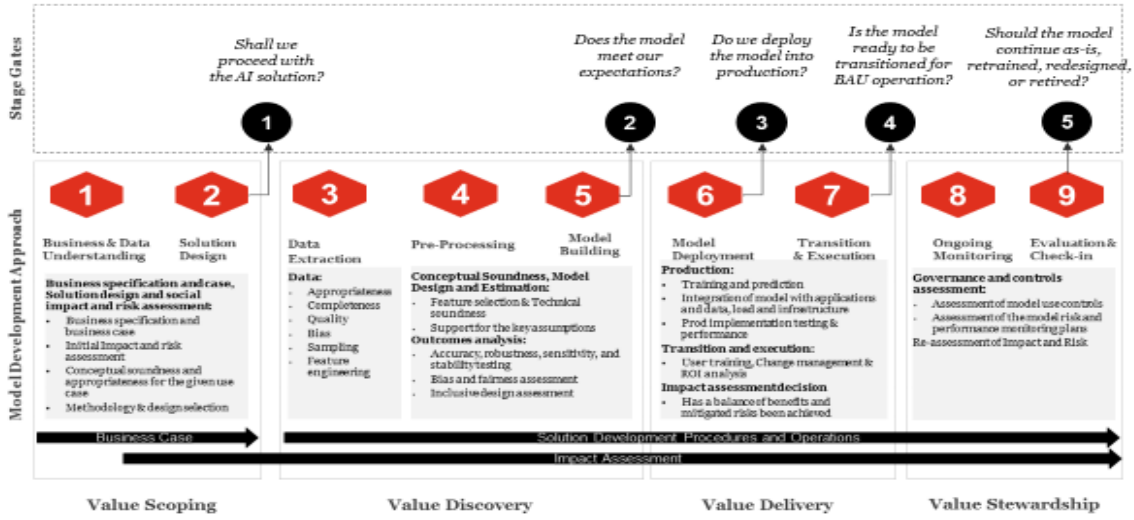
The type of assessment required in the context of an AI application is best served by open ended questions that require a descriptive answer. The assessment of impacts is best done by involving a broad cross section of internal stakeholders and in some cases with the advice of external representation, such as an ethics or data review board. Most organizations likely will incorporate this model into other existing business processes and adapt it for their own needs. This structure for an AIA brings many of these common organizational components together into a risk-based decision-making format. Ultimately, it is up to the organization to determine which elements they include in or add to each section, according to their existing assessments, business models and other compliance and commitment needs.

A key part of an AIA assessment system involves what form of review should happen at what development stage of the AIS. PWC has developed a 9 step AI model life-cycle[10] and from this a five stage gate review to successful AI Governance.[11] Each stage gate review is designed to address the following questions and should involve a broad set of stakeholders and decision makers:

- **Stage Gate 1** - Is it worth having an AI solution or not? How is the AI solution designed (build or buy or rent)?
- **Stage Gate 2** - Does the model meet expectations?
- **Stage Gate 3** - Does the model get deployed into production?
- **Stage Gate 4** - Is the model ready to be transitioned for 'business-as-usual' operation?
- **Stage Gate 5** - Should the model continue as-is, retrained, redesigned, or retired?

The following schematic depicts the key areas to address in each stage gate, including the review stages, which align to specific question in the AIA.

[9] Learn more about AI Ethics Framework | Digital Catapult (migarage.ai)
[10] see HBR Why Most Organizations' Investments in AI Fall Flat
[11] Modified on original Six stage gates to a successful AI governance | by Anand Rao | Towards Data Science



There are several key questions that should form part of each stage gate review. Business case requirements and solution design procedures also should be incorporated. These stage gate reviews should include a review of the iterative evolvement of the AIA. **Appendix II** contains some suggested questions for each stage gate of the review process.

3.1 What is Different about this Approach

Unlike a compliance or technical assessment, an AIA requires an iterative, integrated and holistic approach and review. For example, the impact may not be fully known even at the development stage of AI and the approach to testing may not be evaluated fully until the deployment stage is reached. While an initial assessment of benefits and risks is key at stage 1, it is critical this evaluation be reviewed at stages 4 and 5 because it should form a key part of the go/no-go decision and assessment of mitigating controls. Therefore, an AIA should be thought of as a project document that is iterative but is subject to specific review points along the development lifecycle, including postproduction.

An AIA also is a self-reinforcing governance process. Many questions will require standards and operating procedures to be established as part of an end-to-end governance model, and each AI application will require both an appropriateness and effectiveness evaluation against each of these standards. The flip side is that the establishment of these standards and procedures will be key to informing different organizational roles about what to do as they develop AI solutions. This assessment capability needs to be part of various stages of AI from planning (model scope and purpose), development, deployment, and ongoing model use stages. An AIA must address a full range of governance and impact areas from the technology itself, core data management and issues and the impact of data use; it enables incorporation of ethics into existing compliance exercises without adding much additional burden. Its iterative nature could be very different from a "project sign-off" evaluation that many IT projects are subject to today, requiring an approach different from the way lines of defense may operate. Finally, core to both successful governance and to responsible AI more generally is a sound project management model that encompasses key elements such as change management and operational/resource requirements and management.

Each AI solution an organization considers should be assessed with an AIA as part of end-to-end governance. A comprehensive end-to-end governance framework considers the needs across the organization and throughout the AI and analytics lifecycle. It requires defining comprehensive policies and procedures across the organization. In this model, organization policies have a foundation in ethical values and/or principles, encompass legal and regulatory requirements, and are translated into the language and culture of the organization. End-to-end capabilities require changes to procedures, roles, and responsibilities across a wide range of organizational functions.

4 How Organizations can move Forward

These assessments and governance models need to be adapted and fit into an organization's existing processes and business models as well as their strategy and overall business objectives. Properly designed and implemented, AIAs offer a mechanism to address the complexities of AI solutions, including the ethical and legal aspects and the full range of impacts to all stakeholders. They can help the organization's own goals of mitigating reputational damage and growing the trusted provider status with their stakeholders. AIAs would serve the broader objectives and cover the regulatory aspects of requirements such as DPIAs. Organizations that are adopting "beyond legal compliance" approaches to governance may look to their privacy functions to drive AI governance and assessment oversight as highlighted in the IAF's report on Demonstrable Accountability. However in many cases additional capabilities need to be created[12].

To complement an AIA, there is significant need to address AI governance improvements. To date, proposed solutions or frameworks tend to take two different approaches, geared for different stakeholders: tech-oriented solutions to improve the traceability and lineage of models and data, and process-driven solutions to provide gating mechanisms or standardized templates for documentation. There currently are a multitude of technical solutions in development by software companies, cloud providers, AI/ML platform companies and the open-source community to solve for issues around bias and ethics. While this is a promising development, effective AI governance cannot be a purely technology-led effort as it only solves for concerns technical stakeholders may have; it does nothing to assuage concerns posed by consumers, regulators, and requirements for end-to end governance.

However, the development of a robust governance model that addresses AI risks exposes several potential needs in terms of organizational capability and competency gaps.

- The development and deployment of AI requires technical competency and the right set of tools (e.g., fairness and bias testing). Assessment of the skills and resources needed to meet these needs should be performed with a plan to address these deficiencies.
- Oversight is required to effectively apply business group requirements. This oversight could be performed by separate roles within the business group or by designated support or oversight roles.
- Many business group requirements are codified in practice through sets of standards or operating procedures, often established by second line teams. The creation of these requirements may necessitate adding new resources with newly acquired skills.
- Third line functions such as Internal Audit often rely on established controls. Absent accepted standards, functions will have to develop an initial set of controls against which to assess. A partnership between all business lines should develop and assess control effectiveness iteratively.
- Increasingly, more complex or risky AI may benefit from independent perspectives; some organizations are using Data or Ethical Review Committees to provide advice. Developing the right mix of process and committee member skill is required.

It remains to be seen how regulation will advance relative to an AI impact assessment given the quickly evolving nature and landscape. There likely is reticence for organizations to adopt burdensome governance due to a desire to not over self-regulate and place the organization at a competitive disadvantage. Building off the DPIA may be one way to reduce friction to adoption. This approach provides a useful mechanism to balance the benefits and risks of a given application, and documents critical decisions made at different stages of development, deployment, and use. However, an AIA will be required to achieve demands for the ethical and trustworthy adoption of AI.

END

---

[12] Is Your Privacy Governance Ready for AI? - SPONSOR CONTENT FROM PWC (hbr.org)

Appendix 1 – A Model AIA Example

Introduction - Application Impact Assessment Questions –The assessment is used to ensure that the AI Application impact is identified and managed across the AI lifecycle and that related Ethical Principles have been considered. It is a set of questions used to evaluate the risk an application poses and ensure practices have been considered. Answers provided by the user will be assessed qualitatively. A final approver on go or no-go decisions is needed based on considerations of an appropriate balance of benefits and mitigated risks (which include an assessment of how complete required practices have been followed), that supports the AI application.

Not all AIS would or should require detailed or comprehensive answers to each assessment question; the document serves as a project record and review/approval document that is used and updated at key stage gate checkpoints .

Frequency of Assessments - AI Application Impact Assessments are on-going activities as AI projects progress and are maintained. For a new AI application, the assessments should be conducted from the project onset so that proposed benefits and impact can be identified, and appropriate controls can be considered at an early stage. For an existing system, the assessment should be completed and reviewed periodically (e.g. annually or when there are changes). An AI assessment can only give a snapshot of the risks of the AI systems at a particular time. For high risk AI systems, it is recommended to conduct the AI Application Impact Assessment review more frequently. Key is to establish review gates and to set key "Stage Gate" review points . These can be established based on a project Life Cycle basis or a model development basis.

Core Assessment Questions - The following are a set of "core" questions and an associated impact analysis and decision record. There is a self-repeating nature of the process assessment question and the impact analysis; the level of impact can be directly attributable to the way, manner or method chosen to address the risk process type questions (e.g. testing or data analysis during model development). There are also a key set of questions that apply to each gate review; these are at times repeated based on the different stage of development the model is at. (Suggestion - add people (roles) who should be filling out impact assessment questionnaires)

Legend

- **Text** covers the core questions that are to be addressed in a qualitative manner
- ***Italicized text*** *is added context to support the core question and should be used as an aid to provide the qualitative answer*

Application Assessment Questions	Reference Document
Purpose of AI Activity (AIS) and Accountability	
Provide a brief description of the project - What is the business need/goal/objective for this AI & data activity? What is the defined clear purpose in developing the identified AI solution (e.g. operational efficiency or cost reduction)? *What specific problem will this model solve? Is the problem internal, client specific, or generic to your industry? Is this specific task or problem something that is currently addressed by your team and if so how is it dealt with? What is driving the need for this model? (e.g. Automation, Product/service enhancement, New product or Service, Decision support etc.). What human decision(s) will this AI Solution support/replace/enhance? Have you explored solutions that already exist? If similar solutions exist, why is it*	**Business Case Gate 1**

insufficient for the use case?	
Who are the key stakeholders, and what are the objectives of this initiative for each group of stakeholders (each entity or group of individuals participating and/or being impacted by the data/technology use scenario)? How do these outcomes map to the ethical principles or other external positive justification/validation that is generally accepted? *Objectives should be explicitly describable for each stakeholder and map to some externally validated objective (e.g. broader public policy objective). Consider objectives such as: better/lower cost health care, greater access to health services, or better health outcomes or an improved ability to track and assess health outcomes; more accurate sensors or devices to detect or diagnose health conditions or to improve general wellness; improved education; environmental enhancements such as water conservation, energy cost reduction; infrastructure enhancements; economic improvement; more accessible/usable technology; increased job opportunities; protection of reasonable expectation of privacy, including anonymity; protection of freedom of religion, thought and speech or protection of prohibition against discrimination.* *What is the interest (high-level) for ALL impacted stakeholders? Describe the risks that may be created for each stakeholder. How are expected benefits to outweigh potential risks? What is the acceptance or success criteria for this planned activity? Is there a non-AI solution that is currently in place? How might we solve our user needs? Can AI solve this problem in a unique way? What potential harms associated with this solution have been considered?* *There are many resources available to help teams to think through of the potential unintended harms, you may consider the following resources:* ● *Judgement call https://docs.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/judgmentcall/* ● *Harms model: https://docs.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/harms-modeling/* ● *Framework of harms: https://fpf.org/blog/unfairness-by-algorithm-distilling-the-harms-of-automated-decision-making/* Identify the relevant ethical principles for the solution. How are they to be translated into norms and the design and governance requirements using 'Design for values methodology' *Sample - Who are the stakeholders? Identify and list each group. What is the initial project impact or implication to each of the ethical principle: (1) Fairness, (2) Diversity and Inclusion, (3) Human Agency, (4) Lawfulness and Compliance, (5) Data Privacy, (6) Safety, (7) Accountability, (8) Beneficial AI, (9) 'Cooperation and Openness' and (10) 'Sustainability and Just Transition'*	**Business Case** **Gate 1**
What are the potential risks to each stakeholder? *Explain the high-level potential impact and/or concern the AI project and solution could create and what risks the project could create. If the project triggers any of the risk gating criteria, it will require referral to the more senior leaders (e.g. Risk Management*	**Gate 1** **Business Case**

Committee.	
Based on the risk tiering model, which level is this AIS?	
Have all the governance processes and roles outlined been satisfied and who specifically will be fulfilling each requirement? Describe the plan for the governance model as it relates to the specific model development plan, gate review and the associated timelines, *Roles and responsibilities are needed to operationalize AI and ensure accountability for ethics. Who does what? What is the nature/responsibilities of any governing bodies that will be established for this AIS? Are different teams responsible for model validation and model development? Gat checking timelines might include Data collection; Model tuning; Model testing; Bias and fairness assessment; Deployment, etc. Other steps might include monitoring activities, retesting, deployment, infrastructure readiness and assumptions for long term model improvement (if required), transitioning models from development to pre-production*	**Gate 1** **Business Case**
Performance: Which performance metrics will be used to measure success? How do these metrics align with the business use of the solution? *If this solution is intended to replace an existing process, do you have metrics around the performance of the current process? Use 'If Metric X changes by Y pts or %, then we will move forward with implementation.'* *Other: What other metrics would be taken into consideration to measure the success?*	**Gate 2** **Solution Operating Procedures**
Who has ultimate decision-making authority for the AIS and who else needs to be consulted or needs to review?	
AI Governance Process -	
Planning (AI Development Lifecycle)	
Does the project plan account for each stage of the lifecycle of the AI solution including change management and any organizational changes? *Does the project map to the established Project Management requirements? Are functional and technical requirements fully accounted for & documented? Does a roadmap exist to account for possible future requirements, necessary model updates as required and roles and responsibilities.? Is there a process established to move beyond a PoC? Does your change management plan contain the required organizational cultural shift to accommodate your new AI initiatives?*	**Gate 1 & 2** **Business Case** **Solution Operating Procedures**
How are governance controls to be managed that will enable a consistent, robust, repeatable development/implementation process? How have all project team members' roles been established and communicated? How confident are you that staff and/or contract resources are equipped with the skills and knowledge they need to take on the project responsibility?	
What is the intended application/scope of the AI project? How did you choose the model's suitability for the task at hand? How have the business success criteria for a given solution been defined and aligned to function and non-functional requirements?	

Describe which model results will be used to evaluate the model. How are they related to the business question and success criteria? (Tasks examples: Do you validate that the metrics used to select a model are appropriate? Do you evaluate models for reliability? Do you test for model sensitivity? Is there a process to identify model vulnerabilities?) Have you clearly identified users of the product and can you explain usage patterns that will lead to the user getting expected value from the product? How will AI solve the problem in a unique way? Have you explained why certain metrics will be used to measure success? How were AI performance indicators chosen?	
How has "fairness" been described and what steps are in place to measure and test for achieving this? *Given there is no single definition of fairness that will apply equally well to different AIS applications, the goal is to detect and mitigate fairness-related harms as much as possible. AI systems can behave unfairly due to biases inherent in the data sets used to train them or biases that are explicit or implicitly reflected in decisions made by the development teams or can result in unfair behavior when these systems interact with particular stakeholders after deployment. Types of harm can include allocation, quality of service, stereotyping, denigration, over or underrepresentation. They can also be affected by tradeoffs between expected benefits and potential harms for different stakeholder groups. Consider processes that scrutinize the system vision and what resulting potential fairness-related harms to stakeholder groups. Consider defining and scrutinizing the system architecture for example machine learning models, performance metrics user interfaces, data sets needed to develop and test the system. Scrutinize the production datasets against defined fairness criteria. Consider doing a ship review before launch and a code review. Build in regular product review meetings.*	**Gate 2** **Solution Operating Procedures**
How will traceability be maintained across data, experiments, model versions and usage? How will you capture performance against success criteria?? *This could entail documenting methods used for designing and developing the algorithmic system. These could consist of:* *a) Rule-based AI systems: the method of programming or how the model was built;* *b) Learning-based AI systems; the method of training the algorithm, including which input data was gathered and selected, and how this occurred. Describe the methods used to test and validate the algorithmic system:* *a) Rule-based AI systems; the scenarios or cases used in order to test and validate;* *b) Learning-based model: information about the data used to test and validate. Describe the outcomes of the algorithmic system: The outcomes of or decisions taken by the algorithm, as well as potential other decisions that would result from different cases (for example, for other subgroups of users).*	**Gate 2** **Solution Operating Procedures** **DataSheet**
Ecosystem (AI Development Lifecycle)	
What specific types of and sources of data will be collected, tracked, transferred, used, stored or processed as part of model development or application? Is any of the data to be used personally identifiable or human characteristic? Would any proprietary, customer-specific data be needed/used? Is any data sensitive? Provide a link to the datasheet for the dataset to be used for this solution. *Is there an existing dataset in place that can be used to develop the model/solution? If labels are needed, is the dataset sufficiently annotated? How is the dataset fit for purpose? Is the data identifiable to a person? Is the data anonymous (or de-identified) and what policy, processes and/or technical measures have been used to minimize the re-identification of the data to an individual? Does an up to date data dictionary exist for the data used by the project? Have you considered how combinations of characteristics may be used by the model to reveal a special category, which may result in processing that is unfair to the individuals represented by the*	**Gate 2** **Solution Operating Procedures** **DataSheet**

data? Do you collect and use biometric data in your AI model? Source data could be from other internal sources or from externally obtained data. Will you need to conduct or amend a PIA?	
What is the process to log the data lineage to understand the source, path, license or other obligations and transformations of data being loaded into ML Models? Do you have procedures in place to validate use of the data is compliant with any applicable licenses? *What processes are in place to ensure input data is fit for purpose for AI activities? Do you have procedures in place to validate use of the data or model is compliant with any applicable licenses (if the data or technology is 3rd party provided)?*	**Gate 2** **Solution Operating Procedures** **DataSheet**
How have you determined that your data is accurate enough for the purpose of the model and initiative activity? Is the dataset used credible and from a reliable source? What are your techniques for validating the reliability of source data? *Modelling data needs to be appropriate in terms of data sample size and distributions to ensure the AI model makes meaningful and representative inferences. What steps are being taken to determine the accuracy of source data and if the source data will be accurate enough over time? Has consolidation/transformation impacted the data in such a way that accuracy is affected? Are there concerns about the quality of the final data set relative to the purpose of the activity? Tests may include.* • *Accuracy testing* • *Robustness & sensitivity testing* • *Bias & fairness testing* • *Any other applicable testing performed*	**Gate 2** **Solution Operating Procedures** **DataSheet**
Third-party Questions (AI Development Lifecycle)	
Does your AIS contemplate using 3rd party technology or data as either a supplier or partner. Yes? No? *3rd parties are usually external vendors, providers or partners.*	***If no, skip the 3rd party section.***
If your organization obtained models or datasets from a third party, did your organization assess and manage the risks of using these?	
What is/are the documentation requirements of 3rd parties? Did you ask for and receive detailed documentation? Does the documentation satisfy the requirements established? *The applicability of the organization's data to the vendor model should be assessed. Back testing, model validation, and outcomes analysis should be done on the intended portfolio of model use.*	
What are the processes for third parties (e.g. suppliers, end-users, subjects, distributors/vendors or workers) to report potential vulnerabilities, risks or biases in the AI system?	
Development (AI Development Lifecycle)	
What process have you put in place to ensure the quality and integrity of your data? Describe the data cleaning steps you will be using. Describe how you have determined that the data you process is adequate, relevant and limited to what is necessary to achieve the objectives of the model.	**Gate 2** **Solution Operating Procedure**

Complex data cleaning steps are prone to unintended user errors that are difficult to identify and may lead to erroneous modelling results. What type of data do users expect you to accurately share, measure or collect? Is there a process for discovering and dealing with inconsistencies or errors in the data?	
How has the quality of training data been assessed? Were there enough total training samples? Were the samples well-representative of different social groups based on - race, gender, color, age, income, etc.? *Did you consider diversity and representativeness of users in the data being used?*	**Gate 2** **Solution Operating Procedure**
How did you test the performance of the model? Was the model well-trained and analyzed through different metrics - Precision, Recall, F1Score, Accuracy, etc.? What performance metrics did you consider, how did you perform, and did you consider performance differences by subpopulations, e.g. protected groups. Provide a link to the tests scripts/ results document. *Based on the success and acceptance criteria defined as part of the business case,. Describe which model results were used to evaluate the model. Are they different from Gate 1? If so, why were the changes necessary? How are they related to the business question and success criteria? Describe the decision to move forward or not based on model results. Did the model meet the success criteria? If not, were changes made to success criteria to move forward? Did business impact change as a result?*	**Gate 3** **Solution Operating Procedure**
How have all technology/data security requirements been met? How are you verifying that your data sets have not been compromised or hacked? *Did your risk analysis include whether security or network problems such as cybersecurity hazards could pose safety risks or damage due to unintentional behavior of the AI system? Did you put measures or systems in place to ensure the integrity and resilience of the AI solution against potential attacks? What could bad actors do with this data if they had access to it? What is the worst thing someone could do with this data if it were stolen or leaked?*	**Gate 3** **Solution Operating Procedure**
Could the AI system have adversarial, critical or damaging effects (e.g. to human or societal safety) in case of risks or threats such as design or technical faults, defects, outages, attacks, misuse, inappropriate or malicious use? How have these risks been identified and mitigated? *Did you consider different types of vulnerabilities and potential entry points for attacks such as: Data poisoning (i.e. manipulation of training data); Model evasion (i.e. classifying the data according to the attacker's will); Model inversion (i.e. infer the model parameters). Did you put measures in place to ensure the integrity, robustness and overall security of the AI system against potential attacks over its lifecycle? Did you red team/pen test the system?*	**Gate 3** **Solution Operating Procedure**
Describe the validation processes that will be used? *Are there technical review processes? Is there any independent review?*	**Gate 3** **Solution Operating Procedure**
Can the model and the model output be explained in a simple way? What are the communications plans to achieve explainability to impacted stakeholders? Will the system be able to produce reasons for its decisions or recommendations when required? *Is the way your algorithms work transparently communicated to the people impacted by them? Is there any recourse for people who feel they have been incorrectly or unfairly assessed? (circumstances examples: to non-technical stakeholders; you consider the*	**Gate 4** **Solution Operating Procedure**

purpose and the context under which the explanation is needed; where technical explainability/explicit explanations may not be useful to the audience	
What are the other possible alternatives to the current AI solutions that might be more manual, and how do they perform relative to the AI solution in terms of both accuracy-metrics and (business, legal, economic, social etc.)? Is the manual method less risky than the AI? Is it costlier/high investment?	
Deployment (AI Development Lifecycle)	
Describe how your risk analysis accounts for security or network problems such as cybersecurity and adversarial attacks? *(Risks examples: cause safety risks or damage due to unintentional behavior of the AI system; impact of data leakage; What could bad actors do with this data if they had access to it?)*	**Gate 4** **Solution Operating Procedure**
Are all users treated equally? If not - and your algorithms and predictive technologies prioritize certain information or sets prices or access differently for different users - how would you handle consumer/user demands or government regulations or contractual requirements that require all users be treated equally, or at least transparently unequally?	**Gate 4** **Solution Operating Procedure**
What have you put in place such as a series of steps to monitor, and document the AI system's performance (e.g. accuracy) and acceptance criteria? What steps have been put in place to thoroughly test before deployment to see if the system breaks and if it behaves as intended? What is the rollback plan? Describe the testing performed/intended to be performed as part of the model deployment to ensure the system has been correctly deployed and implemented. *Could a low level of accuracy of the AI system result in critical, adversarial or damaging consequences? Did you put in place measures to ensure that the data (including training data) used to develop the AI system is up to date, of high quality, complete and representative of the environment the system will be deployed in?*	**Gate 4** **Solution Operating Procedure**
Could the AI system cause critical, adversarial, or damaging consequences (e.g. pertaining to human safety) in case of low reliability and/or reproducibility? Is there a well-defined process to monitor if the AI system is meeting the intended goals? Did you test whether specific contexts or conditions need to be taken into account to ensure reproducibility? *Did your organization test the AI model used on different demographic groups to mitigate systematic bias?*	**Gate 4** **Solution Operating Procedure**
Did you have an adequate working definition of "fairness" that you apply in designing AI systems? (review initial Fairness assessment)- Please describe. What is your strategy or a set of procedures to avoid creating or reinforcing unfair bias in the AI system, both regarding the use of input data as well as for the algorithm design? *What are your definitions of unfair bias in your use case? Describe the metrics used to evaluate each of them. Describe the existing best practices for detection, identification and mitigation of unfair biases. What is your risk analysis framework? Describe the risks of unfair bias identified for your use case and the groups described by end-user characteristics for which you evaluated bias.*	**Gate 4** **Solution Operating Procedure**
Describe the processes to test and monitor for potential negative discrimination (bias) during the development, deployment and use phases of the AI system? *Does the AI system potentially negatively discriminate against people? Describe the controls in place to mitigate any detected bias. Is the data used in your processing representative of the population you apply the AI to?*	**Gate 4** **Solution Operating Procedure**
Is the AI system designed to interact, guide or take decisions by human end-users that affect humans or society? Could the AI	**Gate 4**

system generate confusion for some or all end-users or subjects on whether a decision, content, advice or outcome is the result of an algorithmic decision? Describe how end-users or other subjects are adequately made aware that a decision, content, advice or outcome is the result of an algorithmic decision? *Could the AI system generate confusion for some or all end-users or subjects on whether they are interacting with a human or AI system?*	**Solution Operating Procedure**
Will the system be replacing human decisions that require judgement or discretion? What is the automated decision the AIA will make? Will they have a legal or similar impact on an individual? *Processing that aims at taking decisions on data subjects producing "legal effects concerning the individual" or which "similarly significantly affects the natural person". For example, the processing may lead to the exclusion or discrimination against individuals Did you evaluate whether a model requires human intervention?*	**Gate 4** **Solution Operating Procedure**
How have you estimated the likely impact of your AI system when it provides inaccurate results? *Did you verify what harm would be caused if the AI system makes inaccurate predictions?*	**Gate 4** **Solution Operating Procedure**
Describe the steps you have put in place to explain the decision(s) of the AI system to the users *In cases of interactive AI systems (e.g., chatbots, robot-lawyers), do you communicate to users that they are interacting with an AI system instead of a human? Did you establish mechanisms to inform users about the purpose, criteria and limitations of the decision(s) generated by the AI system?*	**Gate 4** **Solution Operating Procedure**
Based on the risk tiering model, which level is this AIS now?	
Operate & Monitor (AI Development Lifecycle)	
Describe the detection and response mechanisms for undesirable adverse effects of the AI system for the end-user or subject *Did you ensure a 'stop button' or procedure to safely abort an operation when needed? Did you take any specific oversight and control measures to reflect the self-learning or autonomous nature of the AI system?*	**Gate 5** **Solution Operating Procedure**
Do you have a process in place to incorporate customer/user feedback? Do you have a process to identify application weaknesses? What is the escalation process to address significant issues that may be identified? *Do you have established processes for third parties (e.g. suppliers, consumers, distributors/vendors) or workers to report potential vulnerabilities, risks or biases in the AI system? Upon AIS deployment, ongoing operational support must be established to ensure that the AIS performance remains consistent, reliable and robust.*	**Gate 5** **Solution Operating Procedure**
How are you monitoring the ongoing performance of your model? What is the business continuity plan you have in place? What are your triggers for model maintenance and roll back? *Did you put in place verification and validation methods and documentation (e.g. logging) to evaluate and ensure different aspects of the AI system's reliability and reproducibility? Did you put in place measures that address the traceability of the AI system during*	**Gate 5** **Solution Operating Procedure**

its entire lifecycle? Did you put in place measures to continuously assess the quality of the input data to the AI system? Did you define tested failsafe fallback plans to address AI system errors of whatever origin and put governance procedures in place to trigger them? Are the ongoing monitoring criteria (I.e., performance, bias, explainability etc.), the thresholds for intervention, and frequency of testing established and agreed?	
Compliance (AI Development Lifecycle)	
After deployment, what is the process to continually identify, review and mitigate risks of using the identified AIS? *Did your organization perform active monitoring, review and regular model tuning when appropriate (e.g. changes to customer behavior, commercial objectives, risks and corporate values)?*	**Gate 5** **Solution Operating Procedure**
Have all key decision points of the AI solution been mapped and do they meet all relevant legislation, internal policy or procedure contract requirements?	**Gate 5** **Solution Operating Procedure**
Impact	
Who are the **stakeholders** who are impacted by this AI solution? *(Stakeholder examples: External stakeholders: users, indirectly affected public, data providers; Internal stakeholders: employees, board of directors, C-suite)*	**Gate 1** **Business Case**
Beneficial Impacts	
What are the benefits of this AIS to the business? Are there benefits for society as a whole? *Determine and describe what the benefits are that could be realized by the organization. Consider factors such as increased revenue, lower costs, improved efficiency, enhanced employee satisfaction, engagement and productivity, enhanced citizen (or workforce) relationship, enhancement or maintenance of brand or reputation, assurance of compliance, fraud prevention, enhancement or maintenance of cyber or physical security, new or improved public services or citizen service, improved manner of marketing, improved ability to assess customer preferences, improvements to innovation or enabling greater, faster, more efficient innovation, improved research processes, improved ability to conduct research and find or enroll study subjects, or improved efficiency with studies, innovative ways to conduct research. Do the benefits of having the model in production outweigh the costs of maintaining it? Are there any social interests served with the deployment of this AIS? How does the project /system contribute to or increase well-being? How will the project /system contribute to human values?*	**Gate 1** **Business Case** **(Review and Update)**
What are the benefits to the defined impacted other stakeholders? Could the AIS be used in a way that may result in a specific stakeholder or group of stakeholders being treated differently in a positive way from other groups of individuals? Can the benefits obtained by various stakeholders be measured? *Determine and describe the positive impacts on the various stakeholders that are expected to come from the application of this technology/data activity. Are there identifiable expectations of individuals, groups of individuals for each beneficial use of the AIS?*	**Gate 1** **Business Case** **(Review and Update)**

Determine what the potential positive goal of the difference in treatment is (if any).	
What are the factors that may limit the realization of any benefits to external stakeholders?	
Negative Impact to Specific Stakeholders	
Considering all the factors relating to the AIS, what are the risks (real and/or perceived) and/or potential adverse impacts to **each** identified stakeholder? [this list in the italics section might need to be better aligned to the PWC Ai environment]. Consider the significance and likelihood of these risks or adverse impacts *Consider the risks or increase in risks (real or perceived) to the identified stakeholder as a result of the application of the AIS. Areas to consider include: perception of technology or data about them being used in an impactful way, an impact to the employee relationship, reduced status and/or well-being; damage to reputation or embarrassment; shock or surprise at the processing activity or the results of the processing; inappropriate discrimination, the possibility of inappropriate access to or misuse of information (e.g. insights or predicative data) by the organization, including sensitive categories of data and directly identifiable data; manipulation of needs or desires/wants of the individual (i.e. creation of a need where one previously did not exist); a negative impact of the technology through a probability-based process, such as a score; Who will have access to information on the AIS and who won't? Will stakeholders who don't have access to this information or data or the insight suffer a setback compared to those who do? What does that setback look like? What new differences will there be between the "haves" and "have-nots" of this information? Would the information use about individuals align with their perception of whether this data/information should be used this way? Determine whether there are other sensitivity issues with the potential use of insights and what aspect of use of potential insights might be considered unfair to the stakeholder. Are all stakeholders treated equally?*	**Gate 1** **Business Case** **(Review and Update)**
Does the AIS potentially negatively discriminate against people on the basis of any of the following grounds (non-exhaustively): sex, race, color, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation?	**Risk Assessment** **Gate 4**
Is it foreseeable that the potential application of AIS or data analytical insights or the data activity might seem surprising, inappropriate or discriminatory or might be considered offensive causing distress or humiliation? *Could the data or technology be used in a way that may result in a group of individuals being treated differently from other groups of individuals?*	**Risk Assessment** **Gate 4**
Are there potential negative impacts of the AI system on the environment? Could the AI system have a negative impact on society at large or democracy? *Did you assess the societal impact of the AI system's use beyond the (end-)user and subject, such as potentially indirectly affected stakeholders or society at large?*	**Risk Assessment** **Gate 4**
For data/technology activities that involve 3rd parties (e.g. receiving or sourcing technology/data as part of this activity), what are the associated risks? *Examples of 3rd parties could include data brokers that sell blocks of information, data aggregators, providers of storage and computing tools, and data trusts. Examples of risks could include data accuracy, data protection, downstream use monitoring and control, legitimate data collection (when done through 3rd parties), data availability.*	**Risk Assessment** **Gate 4**

Is there any likelihood the AIS could lead to any potential costs from the legal and business perspective? *For example, lawsuits can potentially lead to additional legal costs. Is it possible that the AI solution might lead to such overheads?*	**Risk Assessment Gate 4**
Controls	
What are the additional technical and/or procedural safeguards (mitigating controls) that are being implemented to prevent and mitigate risks should they occur? *Have appropriate governance and accountability measures and processes been established? Is the accuracy and/or quality of the data appropriate for the data activity? Does the relative accuracy of the data have an impact on individuals/groups?*	
Describe the mechanism used to externally explain how technology and data is used, how benefits and risks to individuals that are associated with the processing are considered and/or addressed. *Determine what the transparency and individual accountability mechanisms are and whether they are appropriate for the information activity use. Does the application of the technology/information do anything your users don't know about, or would probably be surprised to find out about? What are the explainability mechanisms proposed that are to be used?*	**Gate 2 & DataSheet (Review and Update)**
Describe the extent of human involvement in the data processing. Have all considerations with respect to automated decision making been accounted for? *Have the humans (human-in-the-loop, human-on-the-loop, human-in-command) been given specific training on how to exercise oversight?*	**Gate 2 & DataSheet (Review and Update)**
What is the mechanism to capture feedback by users of the system? Will there be a recourse process planned or established for users of the AIS that wish to challenge the decision? *Could the AIS benefit from additional review and input by an outside party (e.g. review Board)*	**Gate 2 (Review and Update)**
Is the plan and/or process(es) in place to assess the model performance over time sufficient to manage risks, (including model drift and changes in the model use environment to ensure that output stays statistically accurate)? Does it effectively monitor and test to ensure the project's goals, purposes and intended applications are being met as well as the ongoing assessment for fairness and bias?	**Gate 4&5**
Decision – Go /No-Go (Approver)	
How effective are the mechanisms that facilitate the AI system auditability (e.g. traceability of the development process, the sourcing of training data and the logging of the AI system's processes, outcomes, positive and negative impact)? *Could the AI system be audited by independent third parties?*	**Risk Analysis Gate 4&5**
What are the additional requirements surrounding your use of AI? Are there other legal, cross-border, policy, contractual, industry or other obligations linked to the collection, analysis and use(s) of data (or technology)? Have these all been addressed?	**Risk Analysis Gate 4&5**
How effective are the overall controls and safeguards in reducing risk?	**Risk Analysis Gate 4&5**

Decision – Has an appropriate balance of benefits and mitigated risks supports the AIS processing activity been effectively achieved. Does it achieve alignment with ethical principles? Are there any other factors that should be considered? Have the interests, expectations and rights of stakeholders been effectively addressed *Describable, achievable net positive benefit outcomes (tangible benefits to people) have been demonstrated, and negative consequences have been mitigated to a satisfactory and demonstrable level. The proposed uses of technology & data meet all the ethical principles and values of diversity, inclusion, and privacy.* **Does the project or the AIS require escalation to the Senior leaders (e.g. Risk or AI Governance Committee)**	**Risk Analysis & Decision Gate 4&5**

Appendix II – Sample Stage Gate Review Questions

Stage Gate 1- Shall we proceed with the AI solution should first address the **Business Impact;** what business activity, decision, or action are we making more efficient or effective? What is the business impact to the organization that deploys the solution? **Success and Acceptance Criteria** evaluation should address the performance & success criteria of the model; the interpretability, explainability, safety, privacy, security, and robustness of the model needs to be agreed upon. What high-level testing and validation procedures should it meet for success? **Build vs Buy vs Rent Analysis** should cover Does the rented/bought model meet the business requirements, success and acceptance criteria? How is security and confidentiality of data taken into consideration? **Dataset & Model Requirements** should review the process for gathering & collecting the data, collection process, preprocessing, cleansing, and labelling required, recommended uses, sharing and distribution of the data, maintenance, and retiring of data? What are the specifications of the models being built, the intended use of the model, ethical considerations (such as bias and fairness or explainability requirements), training, validation, and test data? The first iteration of the AIA should be reviewed to assess the initial assessment of **Social Impact and Risk Assessment;** Does it infringe on human rights, such as loss of individual liberty, or loss of privacy, or impact the environment adversely? Does it cause physical, emotional, psychological or financial harm? Finally what is the **Development Plan;** What are the specific considerations put in place for development of AI model (AI model development is not an agile or waterfall development)

Stage Gate 2 Does the model meet our expectations and should address the Success **and Acceptance Test Result. For example,** does the model pass all the success and acceptance criteria laid out in the second stage gate? Where did it fail and what was the workaround used? Is there a need to relax the thresholds for acceptance and approve the model for the next stage? **Integration and deployment requirements review** should include will the model inference in production in batch-mode or real-time? How will the model be validated during the transition period and how frequently would the model require retraining? What are the performance tuning and load testing requirements ? How will the model be integrated with the rest of the application systems? There should be a **Change management plan** that addresses what is the plan during any process changes, validation of AI solution, training and redeployment of personnel? **Updated datasheets and model use case requirements** should cover a model architecture that captures how different models interact with each other; how the data is labelled or validated by end users; how the data pipelines are built to gather, cleanse, process and feed the models need to be part of the overall solution design. Is all information related to the solution, its data and model updated? Does it include all changes to the overall model architecture, data pipelines, and solution architecture documented?

Stage Gate 3 - Do we deploy the model into production? Key considerations should address **Integration & deployment test results;** How is the model performing on integration and performance tests? Are performance tests, load balancing , model us bias tests, explainability tests with end users, robustness tests, and adversarial attack tests conducted? For **Model monitoring metrics and requirements,** what are the model monitoring requirements and metrics to be measured and measurement intervals? What is the reasoning for choosing those metrics and measurement intervals? What are the monitoring systems used to measure model performance? A **Process checklist** should address how all changes to processes based on the interaction mode of the deployed model (e.g., human-in-the-loop, human-above-the-loop, human-out-of-the-loop) have been carried out? Is there a detailed process flow for validation of models with appropriate criteria for exceptions, escalations, and approvals? Are ModelOps, DataOps, and SecOps processes in place? For a **Training checklist,** are any new roles (e.g., ModelOps, DataOps etc.) or changes to existing roles documented with clear responsibilities and activities? Is there a requirement for the training of a large number of users (I.e., casual or power users) for new roles or changed roles?

Stage Gate 4 - Is the model ready to be transitioned for 'business-as-usual' operation? Key considerations include **Model Monitoring Results;** Are all data drifts, feature drifts and concept drifts recorded for the model? Have we performed the technical assessment and confirmed that the production model results/thresholds are as observed in the training phase of the model? For the **Business as Usual (BAU) Transition List,** has it been ensured that appropriate operations personnel are assigned and trained? Are the ongoing monitoring criteria (I.e., performance, bias, explainability etc.), the thresholds for intervention, and frequency of testing established and agreed? Do we have the right (dev, data science) team/SMEs in place to manage the model in BAU? A key assessment at this stage is the **Impact analysis** that determines the go/no-

go decision. **Has a balance of benefits and mitigated risks been achieved** ?

Stage Gate 5 - Should the model continue as-is, retrained, redesigned, or retired? Key considerations should address **BAU Monitoring Results;** Is there any decay in the model and how do the results compare between the current period and the previous period/original model deployment and training results? Is there any decay or instability that can result in retraining or redesign of the model? **ROI Value.** What are the benefits of the model (e.g., efficiency gains, effectiveness, enhanced experience, cost savings, revenue gains)? Do the benefits of having the model in production outweigh the costs of maintaining it? **Re-assessment of risk tiering** should be reviewed as well as a **Re-assessment of impact assessment.**

Appendix 3 – Acknowledgements and Additional References

Lead Author

Peter Cullen, Executive Strategist IAF

Peter Cullen is an Executive Strategist at the IAF and President of Global Information Governance Solutions – LinkedIn . He has a business relationship with PWC.

Special thanks to the following individuals who contributed their time and insights

Ilana Golbin
Director, Responsible AI, PWC

Maria Axente
Responsible AI and AI For Good Lead, PWC

Martin Abrams
Executive Director and Chief Strategist, IAF

Barb Lawlor,
Chief Operating Officer & & Senior Foundation Strategist, IAF
Lynn Goldstein
Senior Foundation Strategist, IAF

Top-down and end-to-end governance for the responsible use of AI | by AnandSRao | Towards Data Science
Ten Principles of Responsible AI for Corporates | by AnandSRao | Towards Data Science
Gain trust by addressing the responsible AI gaps | by AnandSRao | Towards Data Science