Background

This document outlines the review questions at the core of AEM's AI practice. Collectively, they provide guidance as to how AI experts, business leaders, and legal representatives can collaborate on decisions about whether and how AI should be used.

This abridged reference is adapted from AEM's internal guidance. In any case, it is often the case that questions must be tailored to new projects. Please reach out to AEM's AI team with any feedback or questions about how governance may be adapted: ai@aemcorp.com.

Introduction

Why is AI governance—and more broadly, responsible AI—so important? Artificial Intelligence (AI) provides a unique set of opportunities and challenges for businesses: while well-designed AI systems have provided dramatic improvement on tasks previously though impossible[1], poorly defined systems have been found to substantially magnify existing issues[2].

The truth is that while successes are easy to measure, failures are a challenge both to define and anticipate, requiring expertise and analysis beyond human supervision of individual decisions. It is critical to follow an AI governance process that allows us not only to adapt to the rapidly evolving legal landscape, but also to avoid unintended harms and enhance public trust.

The goal of AI governance is to provide an actionable and repeatable process that ensures deployed AI systems are legal, safe, within a target risk threshold, and built in such a way that potential negative impacts to individuals, communities, and the environment are minimized. AEM provides documentation and support to ensure that a delivered AI system can be evaluated by the client with respect to their own acceptable risk profile. For this reason, the proposed governance process is designed not only to provide guidance as to appropriate use of AI technologies, but also to build documentation to inform the end-user about the risk and benefits of the AI-enabled system, as well as what strategies have been undertaken to minimize those risks.

The governance process applies to all internal and deliverable projects meeting a formal definition of *Artificial Intelligence* and consists of three parts:

1. The *governance group* oversees formal risk assessment and project evaluation.
2. Although developers and project managers should consult this group whenever they feel it necessary, the process contains a set of *mandatory touchpoints* for low- and high-risk applications where the governance group must be consulted.
3. *Records* are formally kept on all decisions made by the governance group.

[1] Jumper et al., "Highly Accurate Protein Structure Prediction with AlphaFold."
[2] Obermeyer et al., "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations."

Scope

The term Artificial Intelligence was coined to describe computers mimicking human thought or, more accurately, performing actions or making decisions that appear to indicate a human thought process. Federal AI policy is built on this notion, often referencing Section 238(g) of the John S. McCain National Defense Authorization Act for Fiscal Year 2019. While accurately capturing the notion of Artificial Intelligence, the definition—like the term Artificial Intelligence itself—is open to interpretation and therefore not suitable for establishing internal policy.

Instead, AEM currently uses (and recommends) the definition of AI provided by the European Union's AI act, which can be paraphrased as:

> Software that uses one of the following techniques to generate outputs such as content, predictions, recommendations, or decisions:
>
> 1. Machine learning approaches, including supervised, unsupervised, and reinforcement learning, using a wide variety of methods including deep learning.
> 2. Logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference, and deductive engines, (symbolic) reasoning and expert systems.
> 3. Statistical approaches, Bayesian estimation, search, and optimization methods.

All proposed systems that include components meeting these criteria fall under the oversight of the governance process. However, systems that meet the criteria of low risk defined in Appendix A can be exempted from further oversight after a review by the technologist. Additionally, commercially available tools used internally for low-risk applications are not subject to the AI governance process. Instead, it falls under AEM IT policy (Appendix F).

Process

Structure of Governance Group

This has been removed and may be adapted for the unique contexts of other organizations. Generally, AEM recommends a focused group representing diverse perspectives (business, technology, and legal/policy) with a mechanism to engage further with subject-matter experts and affected groups.

Mandatory Touchpoints

Although developers and project managers should seek the guidance of the governance group whenever it is felt necessary, we specify four points in project development when AI-enabled systems must be approved by the governance group.

Condition	Action
An AI-enabled system will be developed…	Before proposal submission, the questionnaire in Appendix B should be submitted. Based on Appendix C, the governance group will respond with an oversight exemption (for low-risk

	applications) or a list of required evaluations and modifications (for high-risk applications).
Prior to use of the new system…	Results of any required evaluations will be reported to the governance group, who will confirm that the results comply with the original need. The Technologist will prepare model and dataset cards.
Following deployment of the system…	Periodic reviews will be scheduled and conducted based on the risk profile of the application.
Any time a change to the scope or method of the AI-enabled system is proposed…	The governance group will re-assess.

Record Keeping

This has been removed and may be adapted for the needs of other organizations.

Revision

New laws and guidance released by American governmental agencies will be immediately reviewed and incorporated into this document and the governance process. The process will additionally be reviewed annually to incorporate experiences of the governance group and may also be updated to reflect laws, policies, or guidance released by foreign governments and commercial entities.

Appendix A: Prohibited and High-Risk Applications

This list of prohibited and high-risk applications are a combination of the EU AI act (Title II (prohibited) and Annex III (high-risk)) and the OMB Proposed Memorandum for the Heads of Executive Departments and Agencies. We additionally add considerations related to the use of sensitive data and unconstrained generative models. Future revisions of this document should integrate additional applications as the governance group is made aware of them, and update with evolving policy.

While the notion of classifying specific prohibited and high-risk applications was chosen to match the approach of the previously noted policies, we recognize that this approach will bias decision-making towards classifying novel applications as low-risk. Both project managers and the governance group are therefore encouraged to exercise discretion when classifying a novel application as high- or low-risk.

Prohibited Applications

1. Applications that deploy subliminal techniques beyond a person's consciousness in order to materially distort a person's behavior in a manner that causes or is likely to cause that person or another person physical or psychological harm.
2. Applications that exploit any of the vulnerabilities of a specific group of persons due to their age, physical or mental disability, in order to materially distort the behavior of a person pertaining to that group in a manner that causes or is likely to cause that person or another person physical or psychological harm.
3. Applications that evaluate or classify trustworthiness based on social behavior or known or predicted personal or personality characteristics, with a social score leading to detrimental or unfavorable treatment unrelated to the context in which the data were generated or collected, or detrimental or unfavorable treatment that is unjustified or disproportionate to their social behavior or its gravity.
4. Real-time remote biometric identification systems in publicly accessible spaces for the purpose of law enforcement, unless it is necessary and used exclusively for:
    a. The targeted search for specific potential victims of crime, including missing children.
    b. The prevention of a specific, substantial, and imminent threat to life or physical safety, or of a terrorist attack.
    c. The detection, localization, identification, or prosecution of a perpetrator or suspect of a criminal offence.

High-risk Applications

1. AI systems for Biometric identification and categorization.
2. AI systems to manage and operate critical infrastructure, such as road traffic, emergency services, and the supply of water, gas, heating, and electricity, among others.
3. AI systems related to education, including making or influence admissions or assignment to educational and vocational institutions, plagiarism detection, monitoring and content filtering, disciplinary intervention detection, determining access to programs, surveillance, and more.
4. AI systems for advertising to, recruiting, or selecting potential employees; or determining promotion, termination, and task allocation of an individual employee.

5. AI systems that control access to and enjoyment of essential private services and public services and benefits such as public assistance, credit, dispatch of emergency services, and ability to acquire housing or insurance and financing for the same. This includes access systems to the same.
6. AI systems for law enforcement, including surveillance, recidivism prediction, offender prediction, and others.
7. AI systems for migration, asylum, and border control management, including support systems such as translation or social media search.
8. AI systems related to enforcement actions for sanctions, trade restrictions, or other controls on exports, investments, or shipping.
9. AI systems for administration of justice and democratic processes, such as those intended to assist a judicial authority in researching and interpreting facts and the law and in applying the law to a concrete set of facts.
10. AI systems that utilize or generate protected information, private information, personally identifying information, information that can cause embarrassment to an individual or group if revealed, or information that can be used in tandem with other information to reveal any of the above.
11. AI systems that control physical movements, including in human-robot teaming, such as the movements of a robotic appendage or body, or the movement of vehicles.
12. AI Systems that apply kinetic force, delivery of biological or chemical agents, or delivery of potentially damaging electromagnetic impulses.
13. AI systems concerning the transport, safety, design, or development of hazardous chemicals, industrial waste, other controlled pollutants, or biological entities or pathways.
14. AI systems related to the design, construction, or testing of industrial equipment, systems, or structures that would pose a meaningful risk to safety if they failed.
15. AI systems for responding to insider threats or accessing or securing government faciltities.
16. AI systems that block, remove, hide, or limit the reach of protected speech.
17. AI systems that detect or measure emotions, thought, or deception in humans.
18. AI systems concerned with medical devices, medical diagnostic tools, clinical diagnosis, as well as mental-health status or violence.
19. AI systems related to child welfare, child custody, or whether a parent or guardian is suitable to gain or retain custody of a child.
20. Public-facing AI systems whose output space is innumerable and, therefore, may contain objectionable or incorrect material (*e.g.*, text, image, or video generation).
    a. *Public-Facing* indicates that individuals not employed by AEM, the client, or an entity directly or indirectly engaged by the client can use the system without supervision.

Appendix B: PM Pre-Development Questions for AI Projects

Project Title:
Technical Point of Contact:
Internal Subject Matter Expert:
Date Submitted:

Introduction
The following questions must be answered by a project manager or relevant technical lead prior to development of a new AI-enabled systems. These questions provide the AI Governance Group with the information required to develop appropriate evaluation procedures and ensure safe operation of developed AI-enabled systems prior to release.

These questions, as well as the related auxiliary information is extracted from the full AEM AI governance document.

Intended Use
Objective: Communicate to the governance group the goal of your proposed AI-enabled system, and what current approaches to compare it to (if any).

- What is the high-level problem being approached by the AI-enabled system?
- What is the current approach to solving this problem (what are you replacing)?
- Does the problem meet the definition of *high-risk* or *prohibited*?
- Are there any potential off label uses meeting the definition of *high-risk* in appendix A? If so, please provide details.
- How will you validate that the system functions as intended?
- What parts of this system are enabled by AI components?
    - What happens if these components return an incorrect answer?

If the problem does not meet the definition of high-risk, you may submit this form now. If you are developing for a high-risk application, please answer the following <u>for each AI component in the system.</u>

For every AI component, please answer the following questions:

Proposed Approach (for every AI component)
Objective: Detail the approach of this specific AI component and its role in the overall system. The definition of AI is provided at the end of the document.

- What is the role of this component in the overall system? Specifically, what is the scope and how does it contribute to the high-risk decision?
- What are the inputs and outputs of this component?
- What alternate approaches (if any) are currently being used to solve this problem?
- What is the planned technical approach? If using a learned method, what is the loss function? If using an LLM, what is the prompt (broadly)?
- If the model fails or is misused, what are the consequences? What guardrails are in place to mitigate these consequences?
- Will performance be validated for this component? If so, how? If not, does evaluation of other components or the full system include evaluation of this component?

Data (for every AI component)

Objective: Provide details on data used to train (if applicable) and evaluate the model. Note that we do not require a full evaluation of the dataset prior to training, but unaddressed dataset issues will be detrimental to the model in ways that may affect final approval.

- What data will be used to train and/or evaluate this method?
    - Are you beginning with a pre-trained model? If so, what dataset was used for pretraining?
    - If procuring a new dataset…
        - Are you collecting the minimal feasible amount of data?
        - Have individuals represented in the dataset consented to use of this data?
            - If yes, are the individuals aware of how data will be used?
            - If no, could use of the data negatively impact these individuals?
        - Will data be shared? How?
    - If using a pre-existing dataset, is this use covered by previously obtained consent?
        - If no, could use of the data negatively impact these individuals?
- Does the data contain features that must be protected or removed? This includes:
    - Personally identifying information, health information, or any other information protected by law.
    - Demographic information such as age, race, gender, or sexual orientation.
- Do the training and/or validation datasets accurately represent the use of the ai-enabled system? Is the distribution likely to shift over time or between use-cases?

Appendix C: Pre-Development Questions for Governance Group

Assessment of Low-Risk Exemption

Objective: Determine whether this application meets the definition of low risk or needs an in-depth review.

- Does this application meet any definition of prohibited or high-risk applications Appendix A? In your judgment, is there another reason this application should be considered high-risk?
- Are there ways this system could be (mis)used that classify as high-risk?

If the application meets the low-risk criteria the review is complete. The point of contact should be advised that if scope changes the application must be re-submitted.

Lessons Learned and Legal Review

Objective: Reference past works and legal documentation to provide context to the proposed work.

- Have any related projects been performed by AEM in the past? What factors were considered? What were the lessons learned?
- Have any related projects been performed or published by other entities? Were any relevant challenges or shortcomings noted?
- What are the legal implications of developing this product?

Overall System

Objective: enumerate the benefits and risks of the overall system to the current method (when used as intended). Failures of the individual AI components and their downstream effects should be addressed within the relevant section.

- What is the current approach to this problem? How do we benchmark the proposed AI-enabled system to the current approach?
- If the system works as intended, what are the benefits?
- If the system works as intended, what are the potential adverse effects?

Summarize: If the system operates as intended, how do the risks and benefits compare to current practice? What quantitative comparison we can make to benchmark against current practice?

Please consider the following <u>for each AI component in the system</u>.

Scope of AI Component

Objective: determine the role, risks, effects, and mitigations of individual AI components.

- What role(s) does this AI component play in the overall system?
    - Can a non-AI component fulfill the same purpose? If so, what quantitative comparison can be performed between the AI and non-AI components?
- If this AI component is taken in isolation, what off-label uses exist?
    - Is this off-label use high-risk?
    - Are there sufficient safeguards in the overall system to prevent this off-label use?
- How may the AI component be inaccurate or fail?
    - If used within scope, how will this affect the behavior of the overall system? What failures have particularly high-impact results?
- Will sufficient guardrails be placed on this AI component?

- Note that human review of outputs is not a sufficient guardrail in many cases, as it will harm overall performance in many cases[3] and prevents underlying issues from being addressed[4].

Summarize: what evaluations need to be run to ensure that the AI component will perform safely and outperform the current approach? What guardrails need to be put into place to ensure that off-label uses or model failures do not introduce significant risks?

Opacity and Alternate Technical Approaches

- Can the output of the AI component be interpreted by the end-user?
- Can the output of the AI component be interpreted by an expert, e.g., for debugging and development?
- If no to either of the above:
  - Is this level of opacity acceptable?
  - Are we able to use a more interpretable approach?
  - Are we able to add post-hoc interpretability to this component?

Summarize: Does this method allow the user sufficient insight into the AI-component's decision-making process? Is the opacity related to features or lead to decisions that may be sensitive? Are there transparent methods or post-hoc transparency strategies that can/should/must be integrated?

Risk Assessment

- If the AI system is used within scope, what are the potential failure modes of this AI component and what are the downstream effects of those failures?
  - Make sure to consider disparate treatment across sensitive attributes.
- If using a data-based method, does the training data reflect the situation that will be seen during deployment? Is the distribution of the deployment data likely to shift over time? Consider differences in language, geography, etc.
  - Can data collection be performed to monitor this drift?

Summarize: Describe the anticipated failure modes of the system and their effect on the downstream task.

Metrics and Measurement

- What quantitative metric(s) can be used to test for the failures and risks described above? For decision systems, consider whether it is best to minimize false-positive rate, miss rate, or overall error.
  - In what ways are these measurements different from the role that the AI is filling?
  - Are there any sensitive features that must be placed under analysis (both individually and in combination)?
- Is the AI component likely to encounter out-of-distribution data? Can this be tested or mitigated? Consider differences in language, geography, etc.
- What is the performance of alternate methods?

[3] Bansal et al., "Does the Whole Exceed Its Parts?"
[4] Green, "The Flaws of Policies Requiring Human Oversight of Government Algorithms."

Summarize: what tests need to be performed to ensure that the system works as intended when placed in an in-label use? What would be considered "success" for these tests?

Summary Report

- Please summarize the risks and benefits of the overall system, including risks posed by failures of individual AI components, and limitations of the approach it replaces.
- Are there any places in which the strategy must be changed to mitigate risk or enhance interpretability?
- What additional guardrails need to be placed on the AI components or the system as a whole?
- What evaluations need to be run to ensure the system will function as intended? How often do we need to re-run the evaluations to guard against distribution shift? What data needs to be collected in situ to perform evaluation against distribution shift?
- What information and fallbacks must be provided to the end-user to ensure that they are properly informed and (if applicable) a non-automated fallback exists?
- Given the risks, benefits compared to the current methods, and proposed mitigations, is this a project that AEM is willing to support? Do we need to develop a contingency plan for system failures?

Appendix D: Full System Card

<project title>

A governance review has determined that this application falls under AEM's definition of *high-risk,* which is based upon definitions provided within US and EU regulations, for the following reason(s):

<item(s) from Appendix A>

To prevent negative outcomes of this high-risk application, we are providing a comprehensive system card that enumerates components meeting our definition of AI, their role in the overall system, acceptable use(s) of both the system and its individual components, and validations and mitigations that have been performed to ensure safety.

If you would like to use this system outside of its approved domain, please contact us at aigovernance@aemcorp.com.

Overall System

- What uses has this system been designed, validated, and approved for?
- What strategy is this system designed to replace?
- If the system works as intended, what are the benefits?
- What are the risks of the overall system, both if the stated criteria are met, or if one or more AI components fail?
- What system components make decisions in a way that cannot be easily understood by a human?

AI Components Summary

The following components in this system meet our definition of AI:

1)

AI Component "X" …

- Intended use:
    - What is the role of this AI component in the overall process? What are its inputs and outputs?
    - What uses has this AI component been validated for?
    - What ways can this component be misused, and what safeguards have been put in to protect it from misuse?
- Technical details
    - What is the high-level strategy?
    - What hyperparameters, architectures, training objectives etc. were used?
        - For learned methods, make sure to include the loss and inputs and outputs.
    - What training/validation data was used? Has the dataset been evaluated for imbalance or potential proxy variables?
        - Is any of the data used for training sensitive? What processes were put in place to protect it?
- Risk Identification and validation:

- What are the potential failures, and what are the effects of these potential failures on the downstream application?
    - Remember to consider out-of-distribution inputs.
- What validations have been performed or guardrails have been put in place to ensure that this these failures do not occur (results should be included)?
- Are there any risks that were identified but could not be assessed? How will these be monitored during deployment?

Appendix E: Low-Risk System Card

<Project Title>

A governance review has determined that this application is *low-risk* under AEM's definition, making it exempt from required oversight and validation for the use-cases described below. We note that our assessment of high- and low-risk does not consider whether the model is likely to produce incorrect outputs, only the severity of the consequences of such incorrect outputs.

The system has been deemed low-risk for the following application(s). If you would like to use this system for any other application, please contact aigovernance@aemcorp.com:
<Target Application>

AI components and consequences of failure are described in the following table:

Component	**Purpose**	**Effect of Failure**

Appendix F: Guidance for Commercial Tools

This has been removed and may be adapted for the unique needs of other organizations.