AI Risk Assessment Process

1. Background and Purpose

Identifying and countering the potentially harmful effects of Artificial Intelligence (AI)-enabled systems is one of the University of California AI Council's goals. The Council created this Risk Assessment Guide to aid in assessing the risks associated with the procurement, development, and deployment of AI-enabled systems, including data privacy, bias, security, and ethical risks.

This guide is intended only for AI procured and/or used for administrative purposes, and does not apply to AI used for research or pedagogy. However, any AI deployed at UC should be consistent with UC's responsible AI principles of Appropriateness; Transparency; Accuracy, Reliability and Safety; Fairness and Non-Discrimination; Privacy and Security; Human Values; Shared Benefit and Prosperity; and Accountability.[i] **This guide will not answer whether a User (the UC location unit deploying the system) can adopt an AI-enabled system**—that is a decision for the UC location's governance. This guide identifies risks that should inform the approval decision and, because it is seldom possible or even desirable to eliminate all risk, ways to manage and mitigate those risks.

2. Approval Considerations

Each campus should establish its own approval processes for AI-enabled systems. Generally, the level of governance approving an AI-enabled system should be informed by that system's characteristics. For example, standard UC Terms and Conditions do not permit a supplier to use AI systems with UC institutional information except with prior written consent from the Chancellor or delegee for the applicable location or as explicitly set forth in the statement of work.

Campuses may also wish to require a Chancellor or delegee's approval of systems to be used for highly consequential decisions, or in areas presumed to be rights- or safety-impacting. Conversely, if the purpose of the system is less consequential and the AI will not use UC institutional information, a campus might choose to designate the head of the unit deploying the AI-enabled system as the appropriate individual to approve its use. Regardless of who approves an AI-enabled system's implementation, that approval should be documented as part of the procurement process and retained in compliance with UC's Records Retention Schedule.

3. How to use this Guide

3.1. Who Should Use It

Typically, the unit deploying the system will be this guide's User. However, determining whether a risk is relevant, or a specific factor fully or partially mitigates that risk, is subjective and requires a degree of informed judgement. Thus, an assessment will likely require the involvement of individuals with an in-depth knowledge of the problem the system is being procured to address and those who have some familiarity with AI and its associated risks. Further, because of the variety of elements being assessed, completing a risk assessment will almost certainly be a collaborative process. Please see Appendix A.3 for suggestions about entities or departments that may be able to answer questions about certain risks and aggravating or mitigating factors.

3.2. Determining Risk Level

To assess the risks of an AI-enabled system, Users should consider the elements in the AI Risk Assessment

[i] University of California Presidential Working Group on AI (2021). *Responsible Artificial Intelligence (Recommendations to Guide the University of California's Artificial Intelligence Strategy)*. https://www.ucop.edu/ethics-compliance-audit-services/compliance/uc-ai-working-group-final-report.pdf

Table in Section 4 of this guide. The Table describes AI risks and recommends consideration of factors that aggravate or mitigate those risks. Users should assess these risks using standard risk assessment considerations: 1) the negative impact or magnitude of harm that will occur and 2) the likelihood of occurrence. Each UC location should integrate this guide into its existing risk management programs. For example, a location could direct users to indicate whether there is **no risk (NR), a mitigated risk (MR) or an unmitigated risk (UR)**. Alternatively, a location might direct Users to **assign numerical values of 1 to 5 based on the User's assessment of that risk, adjusted for aggravating and mitigating factors**. Although, Users should be wary of assessing risk based solely on a total score. A high level of risk related to one factor might be offset by a low level of risk assigned to another factor, resulting in a total score that obscures a significant risk. If a User lacks sufficient information to assess an element, or a supplier declines to disclose the information, the User should consider that element a risk.

3.3. Risk Assessment Process and Cadence

The AI risk assessment process's results should inform procurement, development, and deployment decisions. In addition, campuses may wish to require its use on a defined cadence. Users should assess the risks this guide describes iteratively, at different stages of an AI-enabled system's lifecycle, when a model is being considered for a different use or different data, and at regular intervals. For example, it may be useful to reassess these factors when performing recurrent cybersecurity risk assessments. This guide complements, but does not replace, an AI governance structure or protocols. Finally, units should adapt this guide as needed, building on to it and customizing it to best suit their purposes.

3.4. Risk Appetite or Risk Tolerance

The term "AI" is applied to a variety of technologies, some of which are defined and described in this document's Glossary. Performing an initial risk assessment as described in this guide may help identify whether a more in-depth risk assessment is necessary. Each location must determine whether the risks associated with implementing a specific AI-enabled system are acceptable or exceed the location's risk tolerance. For example, a location might identify specific risks or a maximum number of risks that, if not mitigated, prompt further discussion with relevant individuals. This guide does not provide advice on defining risk appetite or risk tolerance. Risk tolerance should be influenced by legal or regulatory requirements and can be highly contextual and use-case specific. Innovation often involves certain risks, and it would be impractical to implement every mitigating factor described below. However, decisions about risk appetite or risk tolerance (i.e., whether to accept, avoid, mitigate, or transfer the risk) should be made by a UC employee who has been authorized to make these decisions.

3.5. Before Using This Guide

Users should refer to UC's guidance on data classification before using this guide and should determine the classification of any UC data used for training the model, input to the model as a query, and output by the model.[ii] Users who are less familiar with AI should consider reviewing this document's Glossary before proceeding. The questions in section 6. of this document can assist a User with obtaining certain pieces of needed information during a procurement process (such as in a Request for Proposal (RFP) or while collecting information for a sole-source justification). Further, some mitigating factors span multiple risks and are not listed in the table. For example, some risks can be mitigated through function-specific training for executive leadership, legal, labor, privacy, program staff, technical experts, and the general workforce.

[ii] University of California Office of the President. *Home > Policies > Classification of Information and IT Resources*. https://security.ucop.edu/policies/institutional-information-and-it-resource-classification.html

4. AI Risk Assessment Table

This table is organized according to the responsible AI principles identified by UC's Presidential Working Group on AI (listed in the first column on the left). The assessment can be conducted in two phases:

1. An initial assessment should address the five risks in the shaded rows.
2. A full assessment should also include the eight risks in the rows not shaded.

For each factor listed below a User should:

1. Determine if the risk is relevant.
2. Review the aggravating factors to identify whether any are present.
3. Identify the mitigating factors that are present and whether any additional factors can be implemented cost-effectively.
4. Determine the outcome for each risk using their location's evaluation rubric (assign a score, indicate that the risk has been mitigated, etc.).

This table references concepts established in certain laws and regulations. However, those terms and requirements are referenced as best practices and used to assess risk. This table is not intended to assess legal compliance: the legal requirements referenced are not exhaustive and are not relevant to certain use-cases. The endnotes reference documents supporting the concepts in this table and the footnotes provide links to key documents, definitions, and examples. Following the table, a glossary defines certain key terms.

	Risk Description	Aggravating Factor(s)	Mitigating Factor(s)
Appropriateness/Shared Benefit and Prosperity:	1. The AI-enabled system will be used in areas of highly consequential decisions that require a large degree of judgement. These include, but are not limited to, admissions and student conduct, security/policing, health care, hiring and termination.[1]	The system could lead to a state in which human life, health, property, or the environment is endangered.[2] Existing sector- or application-specific guidelines and standards, as well as guidelines and standards from fields such as Transportation and Health, can help Identify uses that are safety-impacting.[3] Rights-impacting uses are those that affect civil liberties and civil rights. Specific examples include, life, liberty, and the pursuit of happiness; the rights to vote, due process of law, and privacy; and the freedoms of speech, thought, and assembly.[4] (See link in the footnote to a more extensive list of rights for consideration.) [iii]	If life and liberty are at stake, there is maximum transparency and accountability?[5] Increased transparency, formal explainability and accountability mechanisms, and/or requiring a human to consider the social context, the precise decisions enabled by the AI-enabled system, its limitations, and the variables it uses.[6] Increased breadth and diversity of input from interested parties, including subject matter experts' review of the test, evaluation, verification, and validation process.[7] An individual may opt out of the decision being made by the AI-enabled system and have a human conduct the analysis instead.[8] The system's results are reviewed or validated by a human.[9]

[iii] See examples of rights relevant to algorithm assessment: Impact Assessment - Fundamental Rights and Algorithms." Ministry of the Interior and Kingdom Relations, (2022). https://www.government.nl/binaries/government/documenten/reports/2022/03/31/impact-assessment-fundamental-rights-and-algorithms/Fundamental+Rights+and+Algorithms+Impact+Assessment.pdf.

	Risk Description	Aggravating Factor(s)	Mitigating Factor(s)
	1.a. There are potential downstream impacts associated with using the AI model's output.[10]	The AI-enabled system is used for a purpose identified as potentially infringing on human rights (see Appendix A.1).[11]	Affected stakeholder groups have been consulted, their concerns have been addressed, and the concerns and their resolution have been documented.[12] The deployer of the AI-enabled system has co-created, with affected stakeholder groups, a system to regularly audit outputs and implement changes as needed.[13]
	1.b. Use of the system could cause inadvertent intellectual property (IP) infringement (e.g., the outputs are insufficiently transformative from existing protected works) or the system was trained on IP that it does not have the right to use.[14]	The system generates written, visual, or auditory content. The system has been trained using protected content (e.g. copyrighted and trademarked work) without obtaining the owners' permission.[15] The system can be prompted with direct references to copyrighted and trademarked works.[16] The system can be prompted to create content that mimics an individual's, voice, likeness, or other distinct features.[17]	The AI model provider confirms that its training data was properly licensed or that it only used open-source data.[18] The AI model provider provides indemnity against copyright infringement.[19] Measures have been implemented to check outputs for infringement.[20] The AI has been trained solely on licensed, public domain, or own data.[21] The system incorporates protections to prevent violations of an individual's rights of publicity.
Privacy and Security:	2. UC data will become part of the AI model.	UC data will be used to train or refine and customize the AI model.[22] Data input as queries and/or output data will be retained or incorporated into the model.[23] The data are classified as P3 or P4.[24] The AI model is adaptive or engages in dynamic training.[25] The system allows access to the underlying data used to refine its operations.[26] The system is for widespread or general use.	The data being used are P1 data, the use of P2 data has been minimized, and all P3 and P4 data have been de-identified.[27] The contract terms provide for a UC-only instance of the system that is isolated from the "parent" system or model and does not share data with the parent system. Data are retained for a defined period and confirmation is provided when the data have been excluded or forgotten from the training set. The IT infrastructure controls access to the data (e.g., the system is used by a single department and isolated from other departments, preventing data from being shared with individuals that should not have access to that data).
	2.a. The output generated by the AI-enabled system may include sensitive data.[28]	The output may contain data classified as P3 or P4.[29] Individuals' identities, or previously private information about them, can be inferred from the AI-enabled system's results.[30]	The data are only exposed to individuals who otherwise have access to them and access to the outputs is shielded from those who should not have access to them.[31]
	2.b. The AI-enabled system is based on an AI model developed by a third party.[32]	UC information will be entered into the model that meets the definition of "bulk data" as defined by the Department of Justice and could be accessed by a country of concern.[33]	"Bulk data" will not be used in the model.[34] The system monitors inputs and actors to determine whether the model is being used for something illegal or inappropriate (e.g.

	Risk Description	Aggravating Factor(s)	Mitigating Factor(s)
		There are no Terms of Service.	generate malicious code, obtain legal advice, procure inappropriate images, etc.).[35] Ownership of inputs and outputs is reserved to UC.
	2.c. The AI-enabled system is based on a model developed by UC.	UC could be held liable for the system's outputs. Information used in the model meets the definition of "bulk data" as defined by the Department of Justice and could be accessed by a country of concern.[36] The model will be available to researchers or users outside of the UC. There are no Terms of Service.	Liability concerns have been identified and addressed. The system is used solely for AI development or for institutional research.[37] The system monitors inputs and actors to determine whether the model is being used for something illegal or inappropriate (e.g. generate malicious code, obtain legal advice, procure inappropriate images, etc.)[38] The system imposes limits on what the model can access (e.g., the system cannot be directed to retrieve confidential information because it does not have access to that information)
Transparency:	3. The system is not transparent, obscuring users' understanding of the system's use of AI and the basis for its recommendations and decisions, thus reducing trust and accountability.[39]	The system does not explain why it made a particular prediction or recommendation.[40] Descriptions of the AI-enabled system's operation (i.e. the mechanisms underlying a system's operation) are not explained or explainable, or they are not interpretable (i.e. the reason why the system made a decision is not clear).[41]	Notice will be provided to users and people affected by its use that an AI model was used.[42] Users are notified that the bases for decisions are not disclosed or explained and that they should check outputs for accuracy. The AI-enabled system's methodology or reason for making a decision is identified and explained.[43] Individuals are able to understand AI-based outcomes, ways to challenge them, and meaningful remedies to address any harms caused.[44] Ongoing testing or monitoring confirms the system is functioning as intended.[45] The AI's decisions are attributed to evidence, such as subsets of training data, or it provides citations or other evidence of the provenance of its outputs.[46] The system operates within its knowledge limits, that is, it only operates for the purpose for which it was designed, or only when it reaches a predefined level of confidence in its output.[47]

	Risk Description	Aggravating Factor(s)	Mitigating Factor(s)
Fairness and Non-Discrimination/Human Values:	4. The bias(es) inherent to the AI model's outputs is/are not disclosed or addressed.[48] Note: see National Institute of Standards and Technology (NIST) defined categories of AI bias.[iv]	Training data has not been assessed for selection, omission, or measurement bias.[49] Data has been de-identified or aggregated. These procedures might result in a loss of accuracy or affect decisions about fairness or other values.[50]	A method for managing the risk of bias has been established and assigned to an individual with the appropriate ability and authority.[51] A process is established for the maintenance of histories, audit logs and other information that can be used to review and evaluate possible sources of error, bias, or vulnerability.[52] The AI Model's training data are assessed to ensure that they accurately and verifiably represent the target population to be served by the AI system.[53] A process is established to test provided explanations for calibration with different audiences including operators, end users, decision makers and decision subjects (individuals for whom decisions are being made), and to enable recourse for consequential system decisions that affect end users or subjects.[54]
	4. a. Use of the AI-enabled system may not align with UC's values.	There is a risk that the system's outputs may not align with UC's Standards of Ethical Conduct (see Appendix A.2).[55] The AI-enabled system's use of resources does not align with UC's sustainable practices policy and climate action policy goals.[56]	The AI-enabled system is regularly tested for biases, inequities, or other unintended consequences.[57] The results of these tests are reviewed and accepted by the location's AI oversight body and maintained in compliance with UC's records retention program.[58] The environmental impact and sustainability of the AI-enabled system are assessed and documented.[59]

[iv] Reva Schwartz, Apostol Vassilev, et al., (2022). *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence*, p.8/77 Fig.2. National Institute of Standards and Technology (NIST), U.S. Department of Commerce. https://doi.org/10.6028/NIST.SP.1270

	Risk Description	Aggravating Factor(s)	Mitigating Factor(s)
Accountability/Accuracy, Reliability and Safety:	5. The development process for the AI model—with respect to demonstrating accuracy, reliability, and safety—is not structured or managed.[60]	The provenance of training data has not been maintained.[61] The system is known to hallucinate.[62]	The development process is structured and documented, and the documentation is maintained in compliance with UC's records retention program.[63] The implementation incorporates rigorous simulation, in-domain testing, real-time monitoring, and the ability to quickly shut down or modify misbehaving systems.[64] The AI-enabled system's decisions can be attributed to subsets of training data.[65] It is easy for a user to identify hallucinated output. Relying on hallucinated output has a low impact. Security risks (data poisoning, model exfiltration) have been considered and addressed.[66]
	5.a. System performance and trustworthiness changes and evolves over time, potentially degrading the accuracy and value of the system's outputs.[67]	The purpose for which the system is being used experiences rapid and significant change.[68]	A plan to regularly update and validate the AI model is established.[69] The system's performance is monitored (see performance monitoring methods described in 5.b).[70]
	5.b The system continues to incorporate new information into the model(s) and, consequently, over time the outputs may change.[71]	The data's reliability has not been assessed.[72]	Training data and testing data are segregated.[73] Procedures are implemented to ensure that (1) data are input in a controlled manner, (2) data are complete, accurate, and valid, (3) any inaccurate information is identified, rejected, and corrected for subsequent processing, and (4) the confidentiality of the data is adequately protected.[74] Performance metrics that map to desired outcomes are established and monitored.[75] Acceptable levels of data drift and model drift are defined.[76] Data drift and model drift are monitored.[77] A plan of action is established for addressing data drift and model drift that exceed acceptable levels.

5. Glossary of Key Terms

Artificial Intelligence (AI):
Artificial intelligence is a tool or system that can perform tasks normally performed by a person. Certain AI can recognize images or speech, learn from data, identify patterns, generate written content or make decisions. AI encompasses many kinds of technologies, such as machine learning (or "ML"), where algorithms learn through experience; and generative AI (or "gen AI," like ChatGPT), which generates new content or data based on a question or data given to the gen AI tool. AI also includes using data collected from past and present events to predict the likelihood of specific outcomes. A key characteristic of AI-enabled systems is their capability to infer.[v]

AI-enabled System or AI System:
An AI system is a machine-based system that is capable of influencing the environment by making recommendations, predictions, or decisions for a given set of objectives. It uses machine and/or human-based inputs/data to: i) perceive environments; ii) abstract these perceptions into models; and iii) interpret the models to formulate options for outcomes. AI systems are designed to operate with varying levels of autonomy.[vi]

Bias:
Bias is broader than demographic balance and data representativeness. NIST has identified three major categories of AI bias to be considered and managed: systemic, computational and statistical, and human-cognitive. Each of these can occur in the absence of prejudice, partiality, or discriminatory intent. Systemic bias can be present in AI datasets, the organizational norms, practices, and processes across the AI lifecycle, and the broader society that uses AI-enabled systems. Computational and statistical biases can be present in AI datasets and algorithmic processes, and often stem from systematic errors due to non-representative samples. Human-cognitive biases relate to how an individual or group perceives AI-enabled system information to make a decision or fill in missing information, or how humans think about purposes and functions of an AI-enabled system. Human-cognitive biases are omnipresent in decision-making processes across the AI lifecycle and system use, including the design, implementation, operation, and maintenance of AI.[vii]

Data Drift and Model Drift:
Data drift refers to changes in the statistical properties of the input data in an operational environment, as compared to the training data. Model drift refers to changes in the relationship between the data inputs and the prediction outputs (i.e., AI-enabled systems may encounter new issues and risks as the environment changes over time. This could mean that the AI-enabled system no longer meets the assumptions and limitations of the original design.[viii]). Data and model drifts could result in performance degradation.

Dynamic Training:
Dynamic training refers to a model that is trained online. That is, data is continually entering the system and incorporated into the model through continuous updates, as opposed to a static model that is trained offline and then used for a while before it is updated or changed. Thus, a system using dynamic training can change and evolve over time and might provide different results to the same query depending on when it is posed.

---

[v] European Parliament (2024). *Artificial Intelligence Act (Regulation (EU) 2024/1689)*. https://artificialintelligenceact.eu/ai-act-explorer/;
[vi] OECD (2022), *OECD Framework for the Classification of AI systems*. *OECD Digital Economy Papers*, No. 323, OECD Publishing. https://doi.org/10.1787/cb6d9eca-en;
[vii] National Institute of Standards and Technology (NIST) (2023). *AI RMF Playbook*. U.S. Department of Commerce. https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook;
[viii] NIST (2023). *AI RMF Playbook*, p.127 Measure 2.4. https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook;

Explainability and Interpretability:
Explainability refers to a representation of the mechanisms underlying AI-enabled systems' operation, whereas interpretability refers to the meaning of AI-enabled systems' output in the context of their designed functional purposes.[ix] Together, explainability and interpretability assist those operating or overseeing an AI-enabled system, as well as users of an AI-enabled system, to gain deeper insights into the functionality and trustworthiness of the system, including its outputs. The underlying assumption is that certain perceptions of negative risk stem from a lack of ability to make sense of, or contextualize, system output appropriately. Explainable and interpretable AI-enabled systems offer information that can help end users understand the purposes and potential impact of an AI-enabled system.[x]

Hallucination:
Hallucination refers to a situation where the model generates content that is not factual or accurate. This includes details, facts, or claims that are fictional, misleading, or entirely fabricated.[xi]

Model:
An AI model is a computational representation of all or part of the external environment of an AI-enabled system – encompassing, for example, processes, objects, ideas, people and/or interactions that take place in that environment. AI models use data and/or expert knowledge provided by humans and/or automated tools to represent, describe and interact with real or virtual environments. Core characteristics include technical type, how the model is built (using expert knowledge, machine learning or both) and how the model is used (for what objectives and using what performance measures).[xii]

Query:
A query generally refers to a question or instruction posed to an AI-enabled system in natural language. This is where machines use Natural Language Processing (NLP) to understand the meaning and intent behind your words. This type of query is used in systems like search engines, virtual assistants (like Siri or Alexa), and chatbots. By understanding the query, the AI can generate relevant responses or complete actions as instructed.

Terms of Service:
Terms of Service refers to the legal terms setting forth the nature, scope, and limits of a service and the rules that the service's users must agree to follow.[xiii]

Training:
AI Model Training refers to the process of feeding the AI model data, examining the results, and altering the model output to increase accuracy and efficacy. To do this, the model needs massive amounts of data that capture the full

---

[ix] Explainability of a machine learning model refers to how easy it is to understand the internal logic the model uses to make a prediction. Linear models (such as logistic regression) and small decision trees are on the more explainable end of the spectrum; neural nets and decision forests are on the less explainable end (often referred to as "black-box").

[x] National Institute of Standards and Technology (NIST). (2023). *Artificial Intelligence Risk Management Framework (AI RMF) 1.0.* U.S. Department of Commerce. https://www.nist.gov/itl/ai-risk-management-framework

[xi] Rawte, V., Sheth, A., & Das, A. (n.d.). *A Survey of Hallucination in "Large" Foundation Models.* https://arxiv.org/pdf/2309.05922

[xii] OECD (2022), *OECD Framework for the Classification of AI systems*, p.20. *OECD Digital Economy Papers*, No. 323, OECD Publishing. https://doi.org/10.1787/cb6d9eca-en

[xiii] https://www.merriam-webster.com/

range of incoming data. In essence, it is the foundation of learning, creating the ability to recognize patterns, understand context, and make appropriate decisions.

Transparency:

The Transparency of an AI-enabled system refers to the extent to which information about the system and its outputs is available to individuals interacting with the system—regardless of whether they are even aware that they are doing so. Meaningful transparency provides access to appropriate levels of information based on the stage of the AI lifecycle and tailored to the role or knowledge of individuals interacting with or using the AI-enabled system. By promoting higher levels of understanding, transparency increases confidence in the AI-enabled system. A transparent system is not necessarily an accurate, privacy-enhanced, secure, or fair system.[xiv]

[xiv] NIST (2023). *AI RMF*, p.15 #3.4. https://www.nist.gov/itl/ai-risk-management-framework

6. Questions for Third Parties

The following are questions useful for informing a risk assessment that can be asked of third parties (such as suppliers providing AI models or systems). Users should assess whether the questions are relevant to their use-case and whether to incorporate these questions earlier (such as in an RFP) or later in the procurement process. The questions about the AI are aligned with the risk numbers in the table.

Questions about the Supplier's Organization:

- Describe the AI governance that the Supplier has established. Who does it include?
- What are the Supplier's policies for using AI technology?
- Has the Supplier obtained a third-party assessment of the AI technology? If so, was the assessment based on the NIST AI Risk Management Framework? Does the assessment comply with relevant regulatory requirements? Please share the results.
- Has the Supplier adopted trusted AI principles? (transparency, explainability, etc.) If so, please describe them.

Questions about the AI:

1. Appropriateness/Shared Benefit and Prosperity:
    - What SMEs provided input on the test, evaluation, verification, and validation processes?
    - What stakeholder groups were consulted, what were their concerns, and how were they addressed?
    - Has a system to audit outputs and make changes been co-created with affected stakeholders?
2. Privacy & Security:
    - Will UC data be used to develop new products or models not part of the UC contract? [xv]
        - Will UC data be used to develop products or models that are part of the UC contract?
    - If UC data will be entered into the system or model, what is the data's privacy level?
    - Is the AI model adaptive or does it engage in dynamic training?
    - How does the system monitor inputs and actors to determine whether it is being used for something illegal or inappropriate?
    - How long are UC data retained and what confirmation is provided when they are destroyed?
3. Transparency:
    - Are users and affected stakeholders provided notice that an AI model was used?
    - Is the AI-enabled system's methodology for making decisions identified and explained?
    - Is there ongoing testing and monitoring of the AI-enabled system?
    - Does the system disclose how it reaches its decisions (such as the subsets of training data used)?
4. Fairness and Non-Discrimination, Human Values:
    - What method is used to manage the risk of bias? Who is responsible for that process?
    - Is the system regularly tested for biases, inequities, or other unintended consequences? How often?
5. Accountability and Accuracy, Reliability and Safety:
    - Does the implementation incorporate rigorous simulation, in-domain testing, real-time monitoring, and the ability to quickly shut down or modify the system?
    - What is the plan to regularly update and validate the AI model?
    - How do you assess the reliability of new information incorporated in the model?
    - Are training and testing data segregated?
    - What are the acceptable levels of data drift and model drift?
    - During testing, what percent of the system's responses were false negatives or false positives?
    - How will the supplier address data drift and model drift that exceed acceptable levels?

[xv] If the Supplier intends to use PHI to train or develop new products or models, doing so could violate HIPAA and should be struck from the agreement. Any other uses of identifiable data should be reviewed by Privacy. Any other uses of de-identified data should be reviewed by the campus unit responsible for data governance.

7. Minimum Risk Scenario

If an AI-enabled System's implementation exhibits all of the following characteristics for a specific use-case, further discussion and analysis may not be necessary, depending on the location's established risk tolerance and AI governance structure.

- The AI-enabled system will not be used in areas of highly consequential decisions that require a large degree of judgement.
- No significant downstream impacts associated with using the AI model's output have been identified
- The AI model is UC specific (isolated from a system accessible to others)
- No UC data will be used to train the AI model
- No UC data will be used to refine the AI model
- No input or output data will be retained or incorporated into the model
- The output generated by the AI model are P1 data
- It is transparent to users and people affected by its use that an AI model was used
- The nature of the AI model's training and the methodology behind its recommendations or decisions are identified and explained
- The AI model's use poses minimal risk of biased results
- The AI model's development process is highly structured and managed
- The AI model is updated and validated regularly according to an established plan
- The AI-enabled system is used solely for AI development or institutional research

8. Appendix A

A.1 – AI Uses Identified as Potentially Infringing on Human Rights[xvi]

- Deploys subliminal techniques or materially distorts people's behavior.
- Exploits people's vulnerabilities due to their age, disability, or social or economic situation.
- Creates or expands facial recognition databases.
- Infers emotions.
- Categorizes people based on biometric data.
- Evaluates or classifies people based on social behavior or personal and personality characteristics.
- Predicts the risk that someone will commit a criminal offense.
- Uses biometric identification systems for the purposes of law enforcement.

A.2 - UC's Standards of Ethical Conduct[xvii]

- Fair Dealing - Members of the University community are expected to conduct themselves ethically, honestly and with integrity in all dealings.
- Individual Responsibility and Accountability - Members of the University community are expected to exercise responsibility appropriate to their position and delegated authorities.
- Respect for Others - The University is committed to the principle of treating each community member with respect and dignity.
- Compliance with Applicable Laws and Regulations - University business is to be conducted in conformance with legal requirements, including contractual commitments undertaken by individuals authorized to bind the University to such commitments.
- Compliance with Applicable University Policies, Procedures and Other Forms of Guidance - Members of the University community are expected to transact all University business in conformance with policies and procedures and have an obligation to become familiar with those that bear on their areas of responsibility.
- Conflicts of Interest or Commitment - Employee members of the University community are expected to devote primary professional allegiance to the University and to the mission of teaching, research and public service.
- Ethical Conduct of Research - All members of the University community engaged in research are expected to conduct their research with integrity and intellectual honesty at all times and with appropriate regard for human and animal subjects.
- Records: Confidentiality/Privacy and Access - The University is the custodian of many types of information, including that which is confidential, proprietary and private. Individuals who have access to such information are expected to be familiar and to comply with applicable laws, University policies, directives and agreements pertaining to access, use, protection and disclosure of such information.
- Internal Controls - All members of the University community are responsible for internal controls. Each business unit or department head is specifically responsible for ensuring that internal controls are established, properly documented and maintained for activities within their jurisdiction.
- Use of University Resources - University resources may only be used for activities on behalf of the University.
- Financial Reporting - All University accounting and financial records, tax reports, expense reports, time sheets and effort reports, and other documents including those submitted to government agencies must be accurate, clear and complete.
- Reporting Violations and Protection from Retaliation - Members of the University community are strongly encouraged to report all known or suspected improper governmental activities (IGAs) under the provisions of the

[xvi] European Parliament (2024). *Artificial Intelligence Act (Regulation (EU) 2024/1689)*. *https://artificialintelligenceact.eu/ai-act-explorer/*

[xvii] University of California (2005). *Statement of Ethical Values.* https://www.ucop.edu/ethics-compliance-audit-services/_files/stmt-stds-ethics.pdf

Policy on Reporting and Investigating Allegations of Suspected Improper Governmental Activities (Whistleblower Policy).

A.3 – Sources of Information

The following table lists suggested sources for information about the risks and aggravating and mitigating factors described in Section 4. Two of these sources are especially important: affected stakeholder groups—including, potentially, UC faculty, staff, and students—and the system developer, which could be UC faculty, staff or the supplier from which UC intends to procure the system. However, please remember that authority to approve the procurement and use of an AI system resides with the individual defined by the campus, as discussed in Section 2.

Risk Description	Potential Information Sources
1. The AI-enabled system will be used in areas of highly consequential decisions that require a large degree of judgement (including, but not limited to, admissions and student conduct, security/policing, health care, hiring and termination).	Unit Implementing, Stakeholder Groups
1.a. There are potential downstream impacts associated with using the AI model's output.	Unit Implementing, Stakeholder Groups, System Developer, Business or Data Architecture, Legal, Compliance
1.b. Use of the system could cause inadvertent IP infringement (e.g. the outputs are insufficiently transformative from existing protected works) or the system was trained on IP that it does not have the right to use.	System Developer, Legal
2. UC Data will become part of the AI model.	Unit Implementing, System Developer, Chief Data Officer, IT Security
2.a. The output generated by the AI-enabled system may include sensitive data.	Unit Implementing, System Developer, Privacy, Legal, Chief Data Officer, IT Security
2.b. The AI-enabled system is based on an AI model developed by a third party.	Unit Implementing, System Developer
2.c. The AI-enabled system is based on a model developed by UC.	Unit Implementing, System Developer
3. The system is not transparent, obscuring users' understanding of the system's use of AI and the basis for its recommendations and decisions, thus reducing trust and accountability.	Unit Implementing, System Developer, Privacy
4. The bias(es) inherent to the AI model's outputs is/are not disclosed or addressed.	System Developer, Diversity, Equity and Inclusion, Legal
4. a. Use of the AI-enabled system may not align with UC's values.	Compliance, Executive Leadership, Stakeholder Groups
5. The development process for the AI model—with respect to demonstrating accuracy, reliability, and safety—is not structured or managed.	System Developer
5.a. System performance and trustworthiness changes and evolves over time, potentially degrading the accuracy and value of the system's outputs.	System Developer
5.b The system continues to incorporate new information into the model and, consequently, over time the outputs may change.	System Developer

9. Revision History

March 14, 2025: Technical update to address rescinded guidance on safety- and rights-impacting AI uses referenced in Risk #1's Aggravating Factor(s).

August 13, 2024: AI Risk Assessment Guidance published by the UC AI Council.

References

[1] National Institute of Standards and Technology (NIST). (2023). *Artificial Intelligence Risk Management Framework* (AI RMF) 1.0, p.1. U.S. Department of Commerce. https://www.nist.gov/itl/ai-risk-management-framework; European Parliament (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)* Arti. 35 § 3(a) – "legal effect concerning the natural person or significantly affect the natural person." https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679; NIST (2023). *AI RMF Playbook*, p.8. The U.S. Department of Commerce. https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook

[2] NIST. (2023). *AI RMF,* p.14. https://www.nist.gov/itl/ai-risk-management-framework

[3] Ibid.., p.15, #3.2. https://www.nist.gov/itl/ai-risk-management-framework.

[4] *U.S. Constitution* art. 1, amends. 1, 3, 4, 5, 9, 14, 15, 19, 24, 26. https://www.archives.gov/founding-docs/constitution-transcript, https://www.archives.gov/founding-docs/bill-of-rights-transcript, https://www.archives.gov/founding-docs/amendments-11-27; Griswold v. Connecticut, 381 U.S. 479 (1965); European Parliament (2012) *Charter of Fundamental Rights of the European Union*, Article 10. https://eur-lex.europa.eu/eli/treaty/char_2012/oj/eng.

[5] Ibid.., p.16 #3.4. https://www.nist.gov/itl/ai-risk-management-framework.

[6] OECD (2022), *OECD Framework for the Classification of AI systems* pp.27, 13 2nd paragraph. *OECD Digital Economy Papers*, No. 323, OECD Publishing. https://doi.org/10.1787/cb6d9eca-en; U.S. Government Accountability Office (GAO) (2021). *Artificial Intelligence, An Accountability Framework for Federal Agencies and Other entities*, section 3.9. https://www.gao.gov/products/gao-21-519sp

[7] NIST (2023). *AI RMF,* p.13 2nd paragraph, #13. https://www.nist.gov/itl/ai-risk-management-framework

[8] European Parliament (2016). General Data Protection Regulation, provision 71. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679

[9] GAO (2021). *Artificial Intelligence, An Accountability Framework for Federal Agencies and Other entities*, pp.53-54. https://www.gao.gov/products/gao-21-519sp

[10] NIST (2023). *AI RMF*, p.8 #1.2.3, p.19 #4. https://www.nist.gov/itl/ai-risk-management-framework

[11] European Parliament (2024). *Artificial Intelligence Act (Regulation (EU) 2024/1689)*, Chapter II, Article 5, provision: 1. (inclusive). *https://artificialintelligenceact.eu/ai-act-explorer/*

[12] GAO (2021). *Artificial Intelligence, An Accountability Framework for Federal Agencies and Other entities*, p.25. https://www.gao.gov/products/gao-21-519sp

[13] Ibid.., section 3.8. https://www.gao.gov/products/gao-21-519sp

[14] NIST (2023). *AI RMF,* p.24 Govern 6. https://www.nist.gov/itl/ai-risk-management-framework; Gil Appel, Juliana Neelbauer, et al. (2023). "*Generative AI Has an Intellectual Property Problem.*" Harvard Business Review. https://hbr.org/2023/04/generative-ai-has-an-intellectual-property-problem

[15] NIST (2023). *AI RMF Playbook*, p.84 Map 4.1. https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook; *Generative AI: Navigating Intellectual Property*. Accessed July 16, 2024. World Intellectual Property Organization. https://www.wipo.int/export/sites/www/about-ip/en/frontier_technologies/pdf/generative-ai-factsheet.pdf

[16] *Generative AI: Navigating Intellectual Property*, p.7. Accessed July 16, 2024. World Intellectual Property Organization. https://www.wipo.int/export/sites/www/about-ip/en/frontier_technologies/pdf/generative-ai-factsheet.pdf

[17] Ibid., p.9.

[18] Ibid., p.6.

[19] Ibid.

[20] Ibid., p.7.

[21] Ibid., pp. 6, 8.

[22] NIST (2023). *AI RMF,* p.8 #1.2.4, p.17 #3.6. https://www.nist.gov/itl/ai-risk-management-framework

[23] Ibid.

[24] Ibid., p.8 #1.2.3 2nd paragraph.

[25] NIST (2023). *AI RMF Playbook*, p.11 Govern 1.5. https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook

[26] NIST (2023). *AI RMF*, p.15 #3.3. https://www.nist.gov/itl/ai-risk-management-framework

[27] Ibid., p.8 #1.2.3.
[28] Ibid., p.17 #3.6, p.8 #1.2.4.
[29] Ibid., p.8 #1.2.3 2nd paragraph.
[30] NIST (2023). *AI RMF*, p.17 #3.6. https://www.nist.gov/itl/ai-risk-management-framework
[31] NIST (2023). *AI RMF Playbook*, pp.114-115 Measure 2.10. https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook
[32] NIST (2023). *AI RMF*, p.5 #1.2.1. https://www.nist.gov/itl/ai-risk-management-framework
[33] *National Security Division; Provisions Regarding Access to Americans' Bulk Sensitive Personal Data and Government-Related Data by Countries of Concern*, 89 F.R. 15780 (proposed March 5, 2024). https://www.govinfo.gov/app/details/FR-2024-03-05/2024-04594
[34] Ibid.
[35] Seyyed Ahmad Javadi, Richard Cloete, Jennifer Cobbe, Michelle Seng Ah Lee and Jatinder Singh (2021). *Monitoring Misuse for Accountable 'Artificial Intelligence as a Service'*. In Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES '20), ACM. https://dl.acm.org/doi/pdf/10.1145/3461702.3462566
[36] *National Security Division; Provisions Regarding Access to Americans' Bulk Sensitive Personal Data and Government-Related Data by Countries of Concern*, 89 FR 15780. https://www.govinfo.gov/app/details/FR-2024-03-05/2024-04594
[37] European Parliament (2024). *Artificial Intelligence Act (Regulation (EU) 2024/1689)*, Chapter I, Article 2, provisions 6 & 8. *https://artificialintelligenceact.eu/ai-act-explorer/*
[38] Seyyed Ahmad Javadi, Richard Cloete, Jennifer Cobbe, Michelle Seng Ah Lee and Jatinder Singh (2021). *Monitoring Misuse for Accountable 'Artificial Intelligence as a Service'*. In Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES '20), ACM. https://dl.acm.org/doi/pdf/10.1145/3461702.3462566
[39] NIST (2023). *AI RMF*, p.6 "Inscrutability," p.15 #3.4. https://www.nist.gov/itl/ai-risk-management-framework; University of California Presidential Working Group on AI (2021). *Responsible Artificial Intelligence (Recommendations to Guide the University of California's Artificial Intelligence Strategy)*, p.8, Principle #2. https://www.ucop.edu/ethics-compliance-audit-services/compliance/presidential-working-group-on-artificial-intelligence.html
[40] NIST (2021). *Four Principles of Explainable Artificial Intelligence*, p.1, 3rd paragraph. The U.S. Department of Commerce. https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8312.pdf
[41] NIST (2023). *AI RMF*, p.16, #3.5. https://www.nist.gov/itl/ai-risk-management-framework
[42] OECD (amended 2024). *Recommendation of the Council on Artificial Intelligence*, section 1.3 item ii. OECD Legal Instruments. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449
[43] NIST (2023). *AI RMF*, p.17 #3.5. https://www.nist.gov/itl/ai-risk-management-framework
[44] University of California Presidential Working Group on AI (2021). *Responsible Artificial Intelligence (Recommendations to Guide the University of California's Artificial Intelligence Strategy)*, p.8, Principle #2. https://www.ucop.edu/ethics-compliance-audit-services/compliance/presidential-working-group-on-artificial-intelligence.html; NIST (2023). *AI RMF*, p.31 table 3 Measure 3.3. https://www.nist.gov/itl/ai-risk-management-framework; Przybocki, P. J. P. a. C. H. P. C. F. D. a. B. M. A. (2021). *Four Principles of Explainable Artificial Intelligence (DRAFT)*. NIST. https://www.nist.gov/publications/four-principles-explainable-artificial-intelligence-draft
[45] NIST (2023). *AI RMF*, p.14. https://www.nist.gov/itl/ai-risk-management-framework
[46] Ibid., p.16 #3.4.
[47] Przybocki, P. J. P. a. C. H. P. C. F. D. a. B. M. A. (2021). *Four Principles of Explainable Artificial Intelligence (DRAFT)*. NIST. https://www.nist.gov/publications/four-principles-explainable-artificial-intelligence-draft
[48] NIST (2023). *AI RMF*, p.40, Item #2. https://www.nist.gov/itl/ai-risk-management-framework
[49] Personal Data Protection Commission of Singapore (PDPC) (2020). *Model Artificial Intelligence Governance Framework, Second Edition*, pp.38-39. PDPC. https://iapp.org/resources/article/pdpc-model-ai-governance-framework-second-edition/
[50] European Parliament (2024). *Artificial Intelligence Act (Regulation (EU) 2024/1689)*, Chapter II, Article 5, provision 1. *https://artificialintelligenceact.eu/ai-act-explorer/*
[51] NIST (2023). *AI RMF*, p.30 Measure 2.11. https://www.nist.gov/itl/ai-risk-management-framework
[52] NIST (2023). *AI RMF Playbook*, p.109. https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook

[53] GAO (2021). *Artificial Intelligence, An Accountability Framework for Federal Agencies and Other entities*, Section 2.2. https://www.gao.gov/products/gao-21-519sp
[54] NIST (2023). *AI RMF Playbook*, p.109. https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook
[55] University of California (2005). *Statement of Ethical Values.* https://www.ucop.edu/ethics-compliance-audit-services/_files/stmt-stds-ethics.pdf
[56] University of California Office of the President (2024). *University of California - Policy on Sustainable Practices*. https://policy.ucop.edu/doc/3100155/SustainablePractices; University of California Office of the President. (n.d.). *UCOP > UC Finance > Capital Programs, Energy and Sustainability > Energy and Sustainability > Sustainability > Policy Areas > Climate Action*. https://www.ucop.edu/sustainability/policy-areas/climate-change/index.html
[57] GAO (2021). *Artificial Intelligence, An Accountability Framework for Federal Agencies and Other entities*, Section 3.8. https://www.gao.gov/products/gao-21-519sp
[58] University of California Office of the President (n.d.). *UCOP > UC Operations > ITS > Initiatives > Records management (UCOP).* https://www.ucop.edu/information-technology-services/initiatives/records-retention-management/index.html
[59] NIST (2023). *AI RMF*, p.30 Measure 2.12. https://www.nist.gov/itl/ai-risk-management-framework
[60] NIST (2023). *AI RMF Playbook*, pp. 85-86 Map 4.2 & pp. 104-106 Measure 2.6. https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook
[61] GAO (2021). *Artificial Intelligence, An Accountability Framework for Federal Agencies and Other entities*, Section 2.1. https://www.gao.gov/products/gao-21-519sp
[62] NIST (2024). *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*, pp.3, 5. U.S. Department of Commerce. https://airc.nist.gov/docs/NIST.AI.600-1.GenAI-Profile.ipd.pdf; Rawte, V., Sheth, A., & Das, A. (n.d.). *A Survey of Hallucination in "Large" Foundation Models*. https://arxiv.org/pdf/2309.05922
[63] NIST (2023). *AI RMF Playbook*, pp. 95-96. https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook; University of California Office of the President (n.d.). *UCOP > UC Operations > ITS > Initiatives > Records management (UCOP).* https://www.ucop.edu/information-technology-services/initiatives/records-retention-management/index.html
[64] NIST (2023). *AI RMF* p.15 #3.2. https://www.nist.gov/itl/ai-risk-management-framework
[65] Ibid., p.16 #3.4.
[66] NIST (2023). *AI RMF Playbook*, p. 107-108 Measure 2.7. https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook
[67] Ibid., p.189 Manage 2.2. https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook
[68] GAO (2021). *Artificial Intelligence, An Accountability Framework for Federal Agencies and Other entities*, Section 4.4. https://www.gao.gov/products/gao-21-519sp
[69] NIST. *AI RMF Playbook*, pp. 102-103 Measure 2.5, pp.53-54 Manage 4.2, p.118. https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook
[70] NIST. (2023). *AI RMF*, p.30. https://www.nist.gov/itl/ai-risk-management-framework
[71] Ibid., p.1.
[72] GAO (2021). *Artificial Intelligence, An Accountability Framework for Federal Agencies and Other entities*, Section 2.2. https://www.gao.gov/products/gao-21-519sp
[73] Ibid., section 2.3.
[74] Ibid., section 2.2; GAO (2009), *Federal Information System Controls Audit Manual (FISCAM)* GAO-09-232G, pp.341-342, https://www.gao.gov/products/gao-09-232g
[75] Ibid., sections 3.5, 3.6, and 3.7.
[76] Ibid., section 4.2; NIST. *AI RMF Playbook*, p.10 Govern 1.4, pp. 40-41 Manage 2.2. https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook
[77] GAO (2021). *Artificial Intelligence, An Accountability Framework for Federal Agencies and Other entities*, p. 64. https://www.gao.gov/products/gao-21-519sp; NIST. *AI RMF Playbook*, pp. 41, 47. https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook