Artificial Intelligence (AI) Risk Management Framework (RMF) Response

Prepared for National Institute of Standards and Technology, Department of Commerce

Submitted August 19th, 2021

modzy®

INTRODUCTION

Modzy, a ModelOps platform and AI marketplace, is pleased to submit a response to NIST's Artificial Intelligence (AI) Risk Management Framework (RMF) RFI. Modzy understands and fully supports NIST's intentions to develop an RMF which helps designers, developers, users and evaluators of AI systems better manage risks across the AI lifecycle. Having recognized the same challenges, Modzy's founders established an AI RMF at the founding of our company. Modzy's AI RMF evaluates and considers potential risks throughout the full AI lifecycle, from design, training, deployment, and maintenance over time. Our multi-disciplinary team and over 20 partner companies have implemented this process to successfully deliver over 100 commercial and open-source AI models across a wide range of model types and potential applications, significantly reducing risk in the process. This framework is both comprehensive and extensible, working on any model type (machine learning, deep learning, computer vision, natural language processing, etc.), and for deployment in even the most secure environments.

We believe our framework, which has been used to deliver mission-critical AI systems to commercial and U.S. government customers, may be of value to NIST. Modzy is a company leading by example—committed to the idea that ethical AI can only be achieved with a robust AI risk management approach. We welcome the opportunity to discuss it further at both the upcoming workshop and in subsequent meetings.

Modzy shares details on our AI RMF approach to NIST to further the development and adoption of AI lifecycle risk management. Our RMF addresses many of the aspects of AI risk management identified by NIST:

1. Modzy's AI RMF helps *identify* AI risks through the Trustworthy-AI Preflight Check which can expose bias in training data or design, the potential for misuse, and the possibility of inadvertent harm being caused by a model.
2. Modzy's AI RMF helps *assess* AI risks through rigorous Model Auditing which helps expose quantitative and qualitative risk factors that are occurring in production, such as sensing adversarial attacks for certain models, or identifying which models are more prone to drift.
3. Modzy's AI RMF helps *respond to* AI risks by embedding adversarial defense into any models during training, embedding human-readable explainability into models, and standardizing models with secure containers.
4. Modzy's AI RMF helps *communicate* AI risks through the Model Transparency Evaluation Rubric and through rigorous documentation which is required for Model Deployment.

The motivation for the Modzy AI RMF started in 2019, when we began scaling the deployment of AI systems in complex environments. The lack of a comprehensive risk management

approach was both slowing the adoption of AI in commercial and government organizations, and letting unknown risks be deployed into production systems. We regularly present and engage in discussion on the Modzy AI RMF in forums such as ISACA Governance, Risk, and Control conference, NVIDIA GTC conference, and publications such as *Data Science Central* and Forbes Technology Council. Among other reports, Gartner® named Modzy a 2021 Representative Vendor for ModelOps and Explainable AI in their 2021 Hype Cycle™ for Data Science and Machine Learning, 2021 Hype Cycle™ for Artificial Intelligence, and 2021 Hype Cycle™ for Analytics and Business Intelligence, and Forrester Research included Modzy as a Provider in their New Tech: Responsible AI Solutions, Q4 2020 report.

RFI Reference: Information provided below is related to the following RFI topics:

6. How current regulatory or regulatory reporting requirements (*e.g.,* local, state, national, international) relate to the use of AI standards, frameworks, models, methodologies, tools, guidelines and best practices, and principles;

7. AI risk management standards, frameworks, models, methodologies, tools, guidelines and best practices, principles, and practices which NIST should consider to ensure that the AI RMF aligns with and supports other efforts;

We are inspired by work from others in this space, including:

- The European Union's Regulatory framework proposal on Artificial Intelligence[i]
- The global landscape of AI ethics guidelines[ii]
- Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims[iii]
- Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI[iv]
- IEEE[v]
- Booz Allen's framework for Assessing the Ethical Risks of Artificial Intelligence[vi]

RFI Reference: Information provided below is related to the following RFI topics:

1. The greatest challenges in improving how AI actors manage AI-related risks—where "manage" means identify, assess, prioritize, respond to, or communicate those risks

AI actors face numerous challenges in managing AI-related risks, all starting with how teams are designing and integrating AI models into production applications. First, there is little rigor applied to the conception and design of AI models. With the democratization of data science and AI, teams are increasingly empowered to access training data and model training tools or frameworks online. This exposes organizations to significant risks related to "shadow AI," or the proliferation of algorithms and AI-enabled systems being developed without the knowledge of a centralized IT or governance body. From there, the process of training models and integrating them into production systems introduces even more problems.

Estimates show that anywhere between 50-90% of models never move from the lab into production systems. Security, broadly, is an afterthought when it comes to AI, leaving models brittle to tampering and susceptible to adversarial attacks to manipulate the outputs by a number of methods. Additionally, many models remain “black boxes,” only understandable to the data scientists involved in the training process.

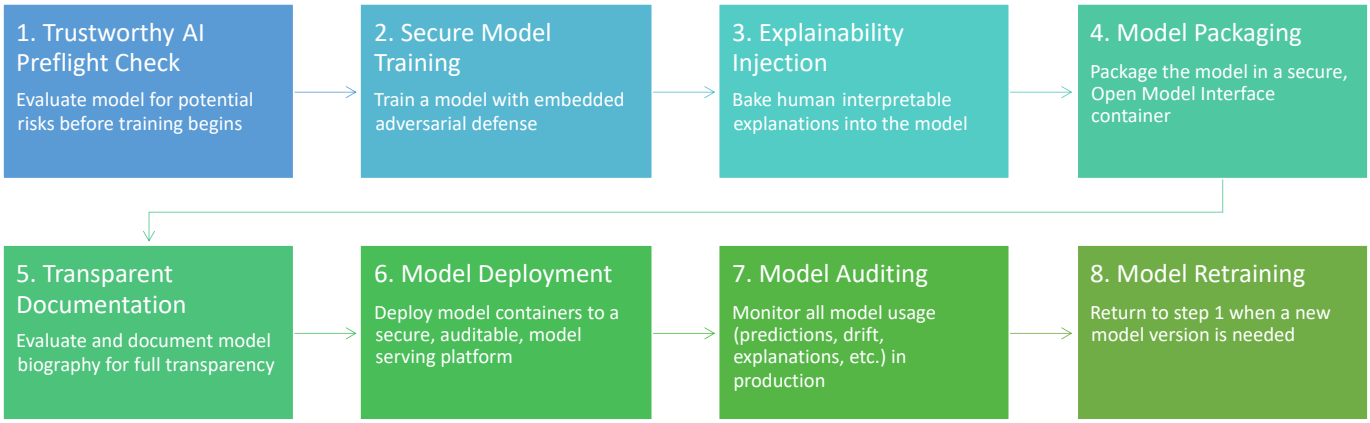
Integrating models into production applications introduces even more challenges. While containerization has become common practice in software development, up until recently there was no standard, interoperable, secure framework for containerizing machine learning models to run at scale, again introducing potential security risks once models are integrated into production systems. From there, the process to upload, deploy, and manage models in use across an organization is haphazard at best. The lack of centralized management tools to monitor AI performance leaves teams flatfooted in attempting to monitor and audit production usage. This leaves teams exposed to even more risk in trying to track important metrics such as model training architecture, training data, expected performance, potential biases, versioning, and more. Without this information in place, it is even more challenging for all teams and stakeholders across an organization to monitor model performance and drift over time.

modzy®

MODZY'S AI RISK MANAGEMENT FRAMEWORK

Our multi-disciplinary team at Modzy, along with more than 20 Partner AI companies, has implemented this process to successfully deliver 100+ commercial and open-source AI models across a wide range of model types and potential applications, significantly reducing risk in the process.

Figure 1: Modzy AI RMF Overall Approach



1: Trustworthy-AI Preflight Check

RFI Reference: Information provided below is related to the following RFI topics:

2. How organizations currently define and manage characteristics of AI trustworthiness and whether there are important characteristics which should be considered in the Framework besides: Accuracy, explainability and interpretability, reliability, privacy, robustness, safety, security (resilience), and mitigation of harmful bias, or harmful outcomes from misuse of the AI;

3. How organizations currently define and manage principles of AI trustworthiness and whether there are important principles which should be considered in the Framework besides: Transparency, fairness, and accountability;

11. How the Framework could be developed to advance the recruitment, hiring, development, and retention of a knowledgeable and skilled workforce necessary to perform AI-related functions within organizations.

Before data is collected or models are trained, models must be evaluated against the Trustworthy AI checklist against five key areas, and a total of 35 sub-areas, which are fully documented in the Appendix:

i. Technical robustness and safety

ii. Privacy and data governance
iii. Transparency
iv. Diversity, non-discrimination, and fairness
v. Accountability

This step ensures that potential adverse impacts of model development are evaluated, documented, and approved prior to model build.

Step 1 in Modzy's AI RMF is dependent on inclusive AI design with a diverse group of stakeholders involved in process. In our experience, inclusive AI design generally produces better performing and more robust algorithms, which is reason enough for most organizations to want to adopt inclusive AI practices. To develop AI models in an inclusive way, organizations must invest in a diverse team, evaluate potential model impact on society before development begins, and develop a list of biases to evaluate against within both training data sets and training approaches. Step 1 in Modzy's RMF ensures this process take place before model development can begin.

Modzy's RMF divides the responsibilities involved in developing AI systems across multiple roles, including data scientists, machine learning engineers, DevOps professionals, and applications developers. This separation of functions ensures that the right people are addressing a narrow range of risks, increasing job satisfaction and retention.

2: Secure Model Training

RFI Reference: Information provided below is related to the following RFI topics:

5. Standards, frameworks, models, methodologies, tools, guidelines and best practices, and principles to identify, assess, prioritize, mitigate, or communicate AI risk and whether any currently meet the minimum attributes described above;

Models must be trained to withstand potential adversarial attacks once they are deployed into production systems. By using an adversarial training method, models gain adversarial defense capabilities, ensuring they remain *robust*, and can maintain the desired level of performance, even in unknown environments.

Given a machine learning model cannot be 100% accurate across variance, irreducible errors, data errors, and human biases, we consider a model to be *robust* when it matches these 3 characteristics:

i. Performance (predictive) metrics for predictions fall within the known validation range

ii. Model tolerance to noise, both random and purposeful, using testing datasets with rare, extreme, and targeted noise to observe and document model performance

iii. When possible, on large-scale models, formal verification techniques (typically using optimization approaches)

This combination of performance testing, tolerance training, and formal verification, represents key areas for the implementation of overall secure model training. The area of secure model training continues to be an active area of research and progress in the larger AI community.

3: Explainability Injection

RFI Reference: Information provided below is related to the following RFI topics:

5. Standards, frameworks, models, methodologies, tools, guidelines and best practices, and principles to identify, assess, prioritize, mitigate, or communicate AI risk and whether any currently meet the minimum attributes described above;

Explainability ensures transparency and insight into model decision-making once integrated into a production application. Incorporating explainability into machine learning pipelines not only ensures stakeholders understand the "why" behind model predictions or recommendations, but also creates a feedback loop and opportunity to generate new labeled datasets for model retraining.

Explainability approaches that rely on a large set of hyper-parameters should be avoided, as they can lead to local instability of explanations and can negatively affect the user's experience and trust in the explainable results.

An explainability algorithm should satisfy 2 properties:

i. It must produce human-interpretable explanations,
ii. It must be locally consistent and efficient at model inference time

4: Model Packaging

RFI Reference: Information provided below is related to the following RFI topics:

5. Standards, frameworks, models, methodologies, tools, guidelines and best practices, and principles to identify, assess, prioritize, mitigate, or communicate AI risk and whether any currently meet the minimum attributes described above

Once models are trained, they must be containerized to a secure container prior to being integrated into a production application. The Open Model Interface (OMI) provides the only standard, interoperable, DISA-compliant, and secure framework for containerizing AI models from any model training tool. OMI is an open standard specification for containerizing machine learning models.

Modzy uses chassis.ml, which implements the OMI standard, to convert models from multiple training tools and frameworks into containers that can run (inference) in many different environments, while maintaining model integrity.

Adopting a containerized machine learning approach for model packaging reduces risk and standardizes model monitoring, performance tuning, and monitoring for output drift detection.

5: Transparent Documentation

RFI Reference: Information provided below is related to the following RFI topics:

2. How organizations currently define and manage characteristics of AI trustworthiness and whether there are important characteristics which should be considered in the Framework besides: Accuracy, explainability and interpretability, reliability, privacy, robustness, safety, security (resilience), and mitigation of harmful bias, or harmful outcomes from misuse of the AI;

3. How organizations currently define and manage principles of AI trustworthiness and whether there are important principles which should be considered in the Framework besides: Transparency, fairness, and accountability

Before models are deployed into production, they must be evaluated against the Model Transparency Evaluation Rubric (see Appendix). The rubric ensures that key metrics related to potential bias, explainability, origin, purpose, testing & validation, and training data are assessed and documented as one final checkpoint before integration into a production application.

The Modzy AI RMF requires validated documentation in the following seven areas:

i. Bias
ii. Explainability
iii. Origin
iv. Purpose
v. Training Data
vi. Model Design
vii. Testing & Validation

Information provided in these seven areas include details on: experiments involved in building the model, system specifications, data set overviews, validation strategy, features and their relative importance, alternative models tested, and explainability features.

6: Model Deployment

RFI Reference: Information provided below is related to the following RFI topics:

5. Standards, frameworks, models, methodologies, tools, guidelines and best practices, and principles to identify, assess, prioritize, mitigate, or communicate AI risk and whether any currently meet the minimum attributes described above;

12. The extent to which the Framework should include governance issues, including but not limited to make up of design and development teams, monitoring and evaluation, and grievance and redress.

To reduce risk in the model deployment and serving process, models must be registered in a model serving platform, commonly referred to as a ModelOps platform. A ModelOps platform is necessary to deploy, manage, monitor, govern, and secure models in use in production applications. All model training information, such as model architecture, training framework, training data, expected performance, versioning, etc. must be documented as part of deployment. Centralized model deployment and management is the only way to mitigate the risks of "shadow AI," or the proliferation of models and tools outside of the purview of a governance or oversight body.

Gartner defines ModelOps, or Model Operations, as the end-to-end governance and life cycle management of all analytics, AI and decision models (including analytical models and models based on machine learning, knowledge graphs, rules, optimization, linguistics, agents and others), allows organizations to deploy models to an auditable AI engine, and audit production usage.vii ModelOps takes DevOps best practices to apply same rigor to how ML/AI pipelines are built and managed to ensure more automation and checks are included as part of the model management process. ModelOps is a way to incorporate risk management into operations that assures model/data integrity, as well as validating reliable operations for both.viii

7: Model Auditing

RFI Reference: Information provided below is related to the following RFI topics:

5. Standards, frameworks, models, methodologies, tools, guidelines and best practices, and principles to identify, assess, prioritize, mitigate, or communicate AI risk and whether any currently meet the minimum attributes described above;

12. The extent to which the Framework should include governance issues, including but not limited to make up of design and development teams, monitoring and evaluation, and grievance and redress.

Deployed model performance must be tracked, monitored, and auditable over time. Teams must establish AI governance, using a centralized management system (ModelOps system) as the basis for holding all teams accountable. A ModelOps tool provides the audit trail and a central location for all AI stakeholders to monitor AI model performance over time. This includes model background information, including model architecture, training data, biases exhibited by the model or within the training data, and expected performance. Teams should be interdisciplinary, and performance monitoring cadence and metrics should be evaluated by use case or application, and potential risks should be triaged by the potential magnitude or scale of impact.

Most factors related to governance, trustworthiness, and risk management are afterthoughts after models are in production. Additionally, many organizations are still in very early stages of establishing an AI governance framework that is key to being able to account for all of these elements; ModelOps is a foundational element to underpinning AI management governance, covered in steps 6 and 7 of Modzy's AI RMF.

Governance is critical to establishing AI trustworthiness and ensuring accountability. Governance and monitoring must align with application or use-case specific concerns. The Framework must establish the need for evaluating the type of impact (e.g. minor annoyance, or harm / death) and scale of impact (e.g. one person, or thousands of people) and establishing the right cadence for governance and monitoring checkpoints.

8: Model Retraining

RFI Reference: Information provided below is related to the following RFI topics:

5. Standards, frameworks, models, methodologies, tools, guidelines and best practices, and principles to identify, assess, prioritize, mitigate, or communicate AI risk and whether any currently meet the minimum attributes described above;

12. The extent to which the Framework should include governance issues, including but not limited to make up of design and development teams, monitoring and evaluation, and grievance and redress.

Models must be monitored for drift over a pre-defined cadence, and retrained as needed. All new model versions should be documented, and tied to their legacy model origins. New model versions follow the entire 8 step process of the Modzy RMF to maintain consistent governance.

MODZY'S AI RMF AS A REFERENCE FOR NIST

Modzy freely offers its AI RMF to be used by NIST as a reference approach to proactively managing risk across the AI lifecycle. Our RMF addresses many of the aspects of AI risk management identified by NIST:

1. Modzy's AI RMF helps *identify* AI risks through the Trustworthy-AI Preflight Check which can expose bias in training data or design, the potential for misuse, and the possibility of inadvertent harm being caused by a model.
2. Modzy's AI RMF helps *assess* AI risks through rigorous Model Auditing which helps expose quantitative and qualitative risk factors that are bearing our in production, such as sensing adversarial attacks for certain models, or identifying which models are more prone to drift.
3. Modzy's AI RMF helps *respond to* AI risks by embedding adversarial defense into any models during training, embedding human-readable explainability into models,

standardizing models with secure containers, and establishing governance and auditing over time

4. Modzy's AI RMF helps *communicate* AI risks through the Model Transparency Evaluation Rubric and through rigorous documentation which is required for Model Deployment.

APPENDIX

Modzy's Trustworthy-AI Preflight Check

Topic	Section	ID	Criteria
Technical robustness and safety	Resilience to attack and security	TR.S.1	Did you assess potential forms of attacks to which the AI system could be vulnerable?
		TR.S.2	Did you put measures or systems in place to ensure the integrity and resilience of the AI system against potential attacks?
		TR.S.3	Did you verify how your system behaves in unexpected situations and environments?
		TR.S.4	Did you consider to what degree your system could be dual-use? If so, did you take suitable preventative measures against this case (including for instance not publishing the research or deploying the system)?
	Accuracy	TR.A.1	Did you assess what level and definition of accuracy would be required in the context of the AI system and use case?
		TR.A.2	Did you verify what harm would be caused if the AI system makes inaccurate predictions?
		TR.A.3	Did you put in place ways to measure whether your system is making an unacceptable amount of inaccurate predictions?
		TR.A.4	Did you put in place a series of steps to increase the system's accuracy?
	Reliability and reproducibility	TR.R.1	Did you put in place a strategy to monitor and test if the AI system is meeting the goals, purposes and intended applications?
Privacy and data governance	Respect for privacy and data Protection	P.R.1	Depending on the use case, did you establish a mechanism allowing others to flag issues related to privacy or data protection in the AI system's

			processes of data collection (for training and operation) and data processing?
		P.R.2	Did you assess the type and scope of data in your data sets (for example whether they contain personal data)?
		P.R.3	Did you consider ways to develop the AI system or train the model without or with minimal use of potentially sensitive or personal data?
		P.R.4	Did you build in mechanisms for notice and control over personal data depending on the use case (such as valid consent and possibility to revoke, when applicable)?
		P.R.5	Did you take measures to enhance privacy, such as via encryption, anonymization, and aggregation?
		P.R.6	Where a Data Privacy Officer (DPO) exists, did you involve this person at an early stage in the process?
	Quality and integrity of data	P.Q.1	Did you align your system with relevant standards (for example ISO, IEEE) or widely adopted protocols for daily data management and governance?
		P.Q.2	Did you establish oversight mechanisms for data collection, storage, processing, and use?
		P.Q.3	Did you assess the extent to which you are in control of the quality of the external data sources used?
		P.Q.4	Did you put in place processes to ensure the quality and integrity of your data? Did you consider other processes? How are you verifying that your data sets have not been compromised or hacked?
	Access to data	P.A.1	What protocols, processes and procedures did you follow to manage and ensure proper data governance?
Transparency	Traceability	T.T.1	Did you establish measures that can ensure traceability?
	Explainability	T.E.1	Did you assess: -to what extent the decisions and hence the outcome made by the AI system can be understood?

			-to what degree the system's decision influences the organization's decision-making processes? -why this particular system was deployed in this specific area? -what the system's business model is (for example, how does it create value for the organization)?
		T.E.2	Did you ensure an explanation as to why the system took a certain choice resulting in a certain outcome that all users can understand?
		T.E.3	Did you design the AI system with interpretability in mind from the start?
Diversity, non-discrimination, and fairness	Unfair bias avoidance	D.U.1	Did you establish a 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?
		D.U.2	Depending on the use case, did you ensure a mechanism that allows others to flag issues related to bias, discrimination, or poor performance of the AI system?
		D.U.3	Did you assess whether there is any possible decision variability that can occur under the same conditions?
		D.U.4	Did you ensure an adequate working definition of "fairness" that you apply in designing AI systems?
	Accessibility and universal design	D.A.1	Did you take the impact of your AI system on the potential user audience into account?
Accountability	Auditability	A.A.1	Did you establish mechanisms that facilitate the system's auditability, such as ensuring traceability and logging of the AI system's processes and outcomes?
		A.A.2	Did you ensure, in applications affecting fundamental rights (including safety-critical applications) that the AI system can be audited independently?
	Documenting trade-offs	A.D.1	Did you establish a mechanism to identify relevant interests and values implicated by the AI system and potential trade-offs between them?

		A.D.2	How do you decide on such trade-offs? Did you ensure that the trade-off decision was documented?
	Ability to redress	A.R.1	Did you establish an adequate set of mechanisms that allows for redress in case of the occurrence of any harm or adverse impact?
		A.R.2	Did you put mechanisms in place both to provide information to (end-)users/third parties about opportunities for redress?

Modzy's Model Transparency Evaluation Rubric

Category	**Question**
Bias	Has this model been evaluated for bias, and has this bias been documented?
Explainability	Does the model provide human-interpretable explanations of how each prediction or decision was reached?
Origin	Does the model include a POC for questions?
Origin	Does the model have a version date?
Origin	Is there a primary developer documented for the model?
Origin	Does the model include clear licensing?
Origin	Does the model have a version number?
Origin	Does the model include a version history?
Purpose	Is there a documented set of intended users for this model?
Purpose	Has the model identified out-of-scope use cases?
Purpose	Does the model provide a list of all possible outputs and explanations of each?
Purpose	Does the model define the types of hardware that can be used to run this model?
Purpose	Is there a documented set of intended uses for this model?
Training Data	Does the model include references to the training data set?
Training Data	Is the full training data set available for access?
Training Data	Does the training data include clear licensing information?
Training Data	Is the training data free and open source (FOSS)?
Training Data	Is there documentation on how the training data set was prepared?
Training Data	Is the training data hashed?
Design	Does the model include technical references that were referred to during the development of this model (papers, conference proceedings, etc.)?
Design	Does the model describe the architecture used to create it?

Design	Does the model provide a list of all dependencies and underlying code?
Design	Do all dependencies (open source or proprietary) of this model include clear licensing information?
Design	Is the model code free and open source (FOSS)?
Testing & Validation	Has the model undergone validation via independent 3rd party performance testing?
Testing & Validation	Is there documentation on how the test data set was prepared?
Testing & Validation	Does the model document it's performance against standard machine learning performance metrics (accuracy, precision, recall, etc.)?
Testing & Validation	Has the model been tested against benchmark data sets?

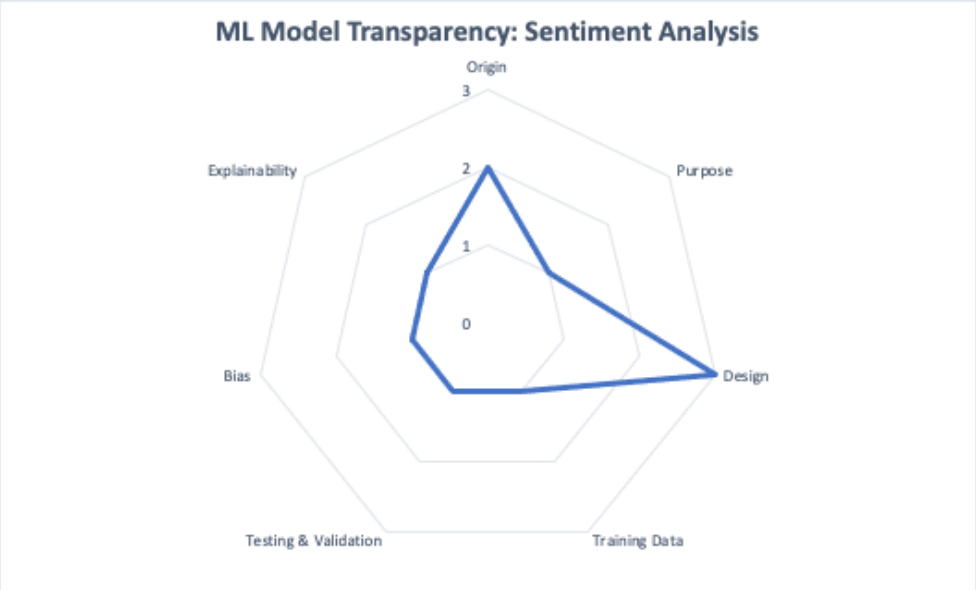
Model Transparency Scores

- Unscored: No transparency form has been completed
- Basic Transparency: A transparency form has been completed but less then 50% of questions have been answered “Yes”
- Moderate Transparency: A transparency form has been completed and 50% or more of questions have been answered "Yes"
- High Transparency: A transparency form has been completed and 75% or more of questions have been answered "Yes"

Example Model Transparency Evaluation

Model Name: Sentiment Analysis

Overall Transparency Score: Basic Transparency



[i] https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai
[ii] https://www.nature.com/articles/s42256-019-0088-2#citeas
[iii] https://arxiv.org/pdf/2004.07213.pdf
[iv] http://www.jennwv.com/papers/checklists.pdf
[v] https://standards.ieee.org/initiatives/artificial-intelligence-systems/index.html
[vi] https://www.boozallen.com/s/insight/publication/assessing-ethical-risks-of-artificial-intelligence.html
[vii] https://www.modzy.com/reports/gartner-hype-cycle-for-data-science-machine-learning/
[viii] https://www.modzy.com/reports/gartner-hype-cycle-for-data-science-machine-learning/