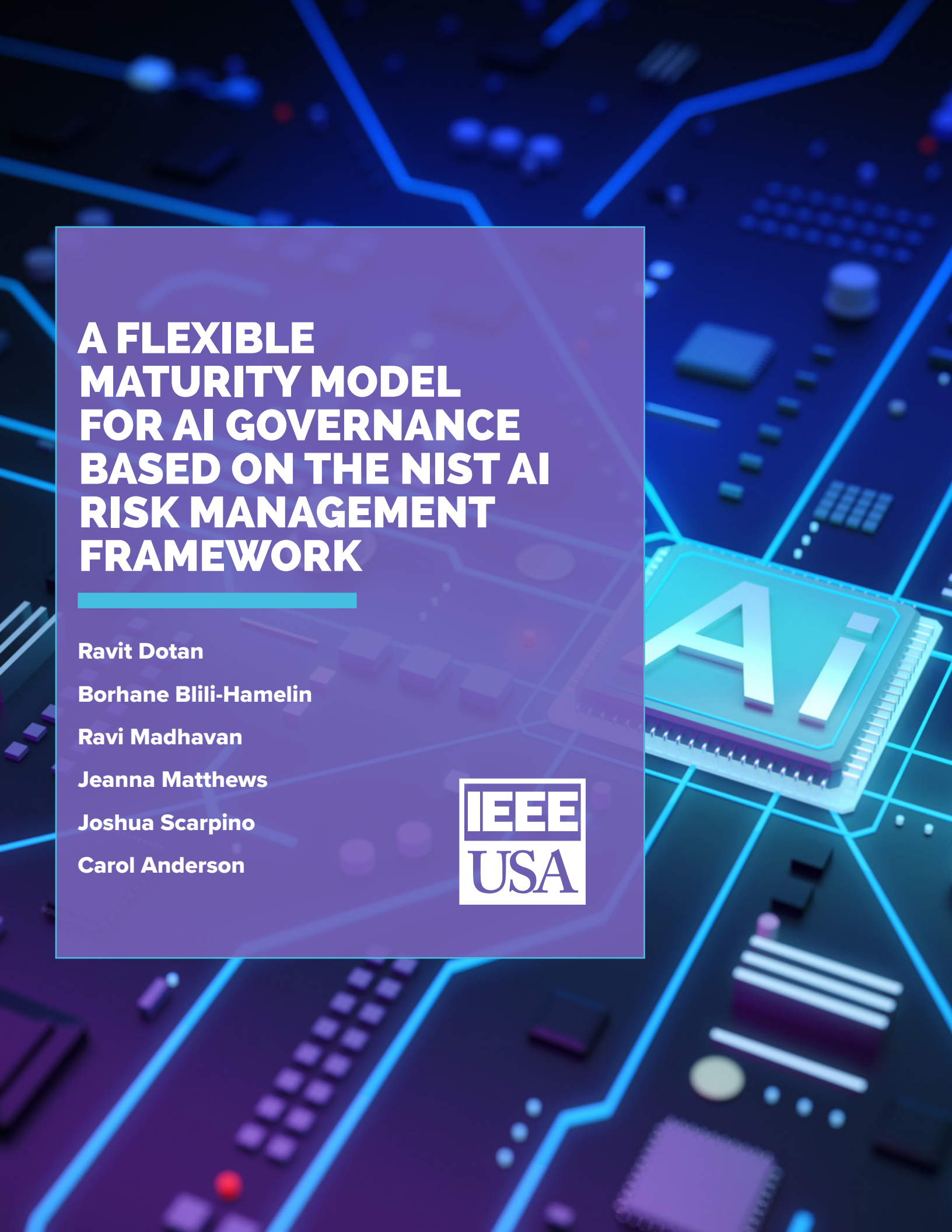


A FLEXIBLE MATURITY MODEL FOR AI GOVERNANCE BASED ON THE NIST AI RISK MANAGEMENT FRAMEWORK

Ravit Dotan

Borhane Blili-Hamelin

Ravi Madhavan

Jeanna Matthews

Joshua Scarpino

Carol Anderson

IEEE USA

A FLEXIBLE MATURITY MODEL FOR AI GOVERNANCE BASED ON THE NIST AI RISK MANAGEMENT FRAMEWORK

Abstract

Researchers, government bodies, and organizations have repeatedly called for a shift in the responsible AI community from general principles to tangible and operationalizable practices in mitigating the potential sociotechnical harms of AI. The AI Risk Management Framework (AI RMF) from the National Institute of Standards and Technology (NIST) embodies an emerging consensus on recommended practices in operationalizing sociotechnical harm mitigation. This paper provides a framework for evaluating where organizations sit relative to the NIST AI RMF.

1. INTRODUCTION

In recent years, increasingly more professionals in the AI ethics space have been calling for "operationalizing AI ethics" or "translating principles into practice" – meaning moving away from articulating general priorities and principles, which has been prominent in the last decade, into establishing processes that rigorously anticipate, evaluate, mitigate, and provide redress for AI harm [1], [2], [3], [4], [5], [6]. Against that backdrop, practitioners, researchers, and government bodies have developed recommendations and practices to bridge the gap [7], [8], [9], [10].

This paper presents a maturity model aiming to help companies decrease this gap. A staple of current technology management toolkits, maturity models are "conceptual multistage models that describe typical patterns in the development of organizational capabilities" [14]. They have been characterized as a Crawl/Walk/Run-style set of factors depicting the progression of capabilities while also serving as a tool to benchmark current capabilities and help set goals and priorities for improvement [17]. The practical utility of maturity models stems from their simplicity, conceptual power, and evolutionary orientation, which result in effective managerial guidance on where to invest attention, effort, and other resources in order to build capability in successive stages (see Poeppelbus et al. [14] and [18] for overviews of the large body of literature on maturity models, and Poeppelbuss & Röglinger [14] for a discussion of design principles for maturity models).

The use of maturity models in technology management dates back to the 1980s, with predecessors dating back to the 1960s [19], [20], [21], [22]. As the popularity of maturity models has increased, their application has spread to many capability arenas, from cybersecurity to software development best practices [14]. Well-known examples include the Software Capability Maturity Model [23] and the risk management maturity model [22]. NIST-related maturity models include the NIST cybersecurity maturity model (National Cybersecurity Center of Excellence [24]), and the simple Privacy maturity model in NIST's privacy framework in the form of "Ready, Set, Go" labels [25]. A maturity model grounded in AI responsibility could help organizations evaluate their existing AI risk management practices and plan how to do better [15].

The maturity model for responsible AI governance presented here is based on the NIST AI Risk Management Framework (AI RMF)[10] (See [49] for more discussion of why we chose the NIST AI RMF, some alternatives we considered, and a discussion of other maturity models).

The NIST AI RMF is a voluntary framework describing best practices for AI risk management, including concrete activities for the development and deployment of AI in a socially responsible

way. It is one of the most well-respected documents on AI governance and is growing in influence, especially in light of the United States October 2023 Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence that specifically calls out the NIST AI RMF [38]. The many AI companies based in the United States may view the US-based policies as especially relevant.

The structure of the paper is as follows. We start by describing the questionnaire and the flexibility it offers. Then, we explain the scoring guidelines and aggregation options. Last, in the appendix, we present the full questionnaire. See [49] for more details on the model's background, design choices, limitations, and implementation examples. See [50] for insights from piloting the model and additional implementation examples.

2. FLEXIBLE QUESTIONNAIRE

This maturity model includes a flexible questionnaire and scoring guidelines. The questionnaire consists of a list of statements, and evaluators are asked to rank them using the scoring guidelines discussed below. The statements in the questionnaire center on concrete and verifiable actions, such as conducting certain processes and documenting the outcomes. For example:

”We regularly evaluate bias and fairness issues caused by our AI systems.”

The questionnaire avoids general and abstract statements such as “Our AI systems are fair.” Further, the statements use the plural first pronoun “we” and the active present tense, e.g., “we evaluate.” This is an intentional choice made to emphasize the responsibilities of the companies and people who manage AI.

The statements cover the content of the RMF’s governance recommendations, which are divided into four pillars: MAP - Learning about AI risks and opportunities; MEASURE - Measuring risks and impacts; MANAGE - Implementing practices to mitigate risks and maximize benefits; and GOVERN - Systematizing and organizing activities across the organization. Each of the pillars includes a list of categories and subcategories. For example, one of the categories in the MEASURE pillar is “MEASURE 2: AI systems are evaluated for trustworthy characteristics.” One of the subcategories in this category is “MEASURE 2.11: Fairness and bias – as identified in the MAP function – are evaluated and results are documented.” [10]. In isolation, each statement in the questionnaire covers one or more of the NIST AI RMF subcategories. For example, the statement above covers the subcategory MEASURE 2.11. Jointly, the statements in the questionnaire cover all RMF subcategories (see the Appendix for the full list).

The questionnaire is flexible in that evaluators are not required to evaluate all statements. The questionnaire allows the evaluator to adjust the evaluation to the organization’s specific context in three key ways: 1) level of granularity, 2) life-cycle stage of the AI system, and 3) multiplicity of AI systems within the organization. We elaborate on each of these in the subsections that follow.

Flexibility 1: Granularity

Evaluators who are interested in a fine-grained evaluation can score each of the individual statements in the questionnaire. However, those who are interested in a more coarse-grained evaluation have an alternative, as the individual statements are divided into nine topics. Each topic is represented by one high level sentence that describes the other lower level statements in that topic. For example, one of the topics is

- Topic 4 - ”*Measuring risk*: We measure our potential negative impacts.”

Under this topic, there are finer-grained individual statements, including for example:

- ”We regularly evaluate bias and fairness issues related to our AI systems.”
- ”We regularly evaluate security issues related to our AI systems.”

Those interested in a coarse-grained evaluation can score only the topic statements. However, the individual statements are still taken into account because, as discussed in more detail below, the scoring guidelines instruct evaluators to give higher scores for better coverage of individual statements.

Flexibility 2: Life-cycle stage

A second aspect of flexibility comes from observing that a subset of RMF subcategories only becomes relevant once the AI system has reached a particular development stage. For example, RMF subcategory MANAGE 4.1 is only relevant after the system has been deployed:

> MANAGE 4.1: Post-deployment AI system monitoring plans are implemented, including mechanisms for capturing and evaluating input from users and other relevant AI actors, appeal and override, decommissioning, incident response, recovery, and change management. [24]

For this reason, the questionnaire is divided into phases of the development lifecycle, based on the AI life cycle described in the RMF [10]. We grouped the life cycle into three stages: (1) Planning and design; (2) Data collection and model building, including verifying and validating the system; and (3) Deployment - including deploying, using, operating, and monitoring the system. Each life-cycle stage contains topics and statements appropriate for that stage from multiple RMF pillars. The evaluator only uses the statements suitable for the relevant life-cycle stage. This flexibility explicitly guides evaluators to avoid questions that are not yet relevant to a particular AI system.

Flexibility 3: Multiplicity of AI systems

Organizations may have multiple AI systems, and the questionnaire allows for flexibility in approaching this multiplicity. Evaluators may score each AI system separately and aggregate those to get scores for the organization as a whole. Those interested in a more coarse-grained evaluation may instead score the organization holistically without delving into the details of each individual system.

Putting it together

Putting together all three aspects of flexibility, those interested in the most fine-grained version of the evaluation will score each AI system using the individual statements appropriate to that system's life-cycle stage. Those interested in the most coarse-grained version of the evaluation will score the organization as a whole using only topic statements appropriate to the life-cycle stage of the most advanced AI system that the organization manages.

Figure 1 illustrates the overall structure of the questionnaire, and the full questionnaire is in the Appendix.

Life-cycle stage	Topic	Sub-Statement	Score for Topic				
			Coverage	Robustness	Input Diversity	Overall	Evidence
Planning and on	[...]						
Data collection, model building, and on	**4. Measuring impact** - We measure the potential negative impacts	[...]					
		4.5 Fairness - We regularly evaluate bias and fairness issues related to this AI system					
		4.6 Privacy - We regularly evaluate privacy issues related to this AI system.					
		[...]					
	[...]						
Deployment and on	[...]						

Figure 1. The structure of the questionnaire (While the full questionnaire is available in the Appendix, here we have retained only one example topic with two example sub-statements. The [...] notation is included to suggest all the other material that has been removed.)

3. SCORING GUIDELINES

Leveraging NIST's work, the scoring guidelines ensure evidence-based and well-reasoned evaluations. The goal is to help organizations communicate constructively when evaluating themselves, understand where they stand, and decide what they should do to improve. To achieve this, evaluators assign scores based on three specific metrics, accompanied by an evidence-based explanation for each score. For case studies and examples, see [49] and [50].

Scoring Metrics

Coverage of RMF Subcategories

As discussed above, the questionnaire allows evaluators to evaluate topic statements only rather than all of the individual statements. For example, the evaluator can score the statement "*Measuring Risk*: we measure the potential negative impacts," but not all the statements it contains, such as "We evaluate bias and fairness issues caused by this AI" and "We evaluate security issues caused by this AI." When the evaluator does so, the scoring of the topic statement should reflect coverage of all the individual statements included in that topic. For example, companies that evaluate and document security but not fairness risks satisfy this metric to a degree lower than companies that address both.

For evaluators who score individual statements, the coverage score reflects whether the company engages in relevant activities or not.

Robustness

The word "robustness" refers to the ideals expressed through NIST's "implementation tiers." The implementation tiers are distinctions NIST uses to describe degrees of risk management activities in areas such as privacy and cyber-security ([25], [24]). These tiers represent an increasing degree of rigor and showcase how well an organization has implemented the component under evaluation. There are four tiers:

1. PARTIAL- Activities are ad-hoc, reactive, occasional, or isolated from key organizational activities.
2. RISK INFORMED - Activities occur but they are informal and irregular.
3. REPEATABLE- Activities are formalized into organization-wide policies or systematic practices.
4. ADAPTIVE - Risk management activities can adapt to changes in the landscape and product, including regular reviews and effective contingency processes to respond to failure.

We have extracted six interrelated ideals for the purposes of maturity evaluation. For convenience, we refer to them collectively as “robustness”:

Robustness - The risk management activities are

1. Regular - Performed in a routine manner
2. Systematic - Follow policies that are well-defined and span company-wide
3. Trained Personnel - Performed by people who are properly trained and whose roles in the activities are clearly defined
4. Sufficiently Resourced - Supported by sufficient resources, including budget, time, compute power, and cutting-edge tools
5. Adaptive - Adapting to changes in the landscape and product, including regular reviews and effective contingency processes to respond to failure
6. Cross-functional - Involve all core business units and senior management. They are informed of the outcomes and contribute to decision-making, strategy, and resource allocation related to the activities (core business units include finance, customer support, HR, marketing, sales, etc)

Input diversity

Input diversity means that risk management activities receive input from diverse internal and external stakeholders. A low level of input diversity means that the relevant activities receive input from relatively few kinds of stakeholders. High levels of input diversity mean that the activities receive input from diverse internal and external stakeholders. For example, suppose that a company chooses its fairness metrics in consultation with civil society organizations, surveys of diverse customers administered by the customer success team, and conversations with diverse employees in the company. In that case, the company demonstrates a high level of input diversity with regard to the statement “We evaluate and document bias and fairness issues related to this AI system.”

The input diversity ideal is not highlighted in the NIST implementation tiers for privacy and cybersecurity. However, it is a key aspect of the AI RMF and is important in AI ethics. AI systems often impact masses of end-users and data subjects as well as society at large. Properly understanding, measuring, and managing AI risks requires an in-depth understanding of the potential impacts which, in turn, requires input from a wide range of perspectives.

Scores

For scoring, evaluators determine the degree to which each statement satisfies the three metrics – low, medium, or high. Then, the statement's score is calculated based on the following rubric, resulting in a score on a scale of 1-5, where 1 is the worst and 5 is the best.

- 5: HHH All three metrics are satisfied to a high degree
- 4: HHM Two of the metrics are satisfied to a high degree and one to a medium degree
- 3: HMM, HHL, HML, or MMM One of the following is the case: (1) Two of the metrics are satisfied to a medium degree and one to a high degree; (2) Two of the metrics are satisfied to a high degree and one to a low degree; (3) One metric is satisfied to a high degree, one to a medium degree, and one to a low degree; or (4) all metrics are satisfied to a medium degree.
- 2: MML, MLL, or HLL One of the following is the case: (1) Two of the metrics are satisfied to a medium degree and one to a low degree; (2) One metric is satisfied to a medium degree and two to a low degree; (3) One of the metrics is satisfied to a high degree and two to a low degree.
- 1: LLL All metrics are satisfied to a low degree
- N/A The statement is not applicable

Explanation

Each score should be accompanied by an evidence-based explanation. Evidence includes information about what organizations do, about what they don't do, and reports of lack of evidence. For example, evidence may include describing artifacts that indicate that the company is engaged in the relevant activities. E.g., they may describe which company documents contain the relevant information and how detailed that information is, the outcomes of the activities, and so on. Evidence may also include indications that certain activities are not performed, which may happen, for example, when company documents imply that these activities are outside of the company's current scope. Further, evidence discussions may also include pointing out a lack of evidence. Evaluators can note in their comments a distinction between lack of any evidence and the presence of evidence to the contrary. Last, evaluators provide evidence that a statement is not applicable, which may happen for example, due to the life-cycle stage of the evaluated AI system. For more information and examples of evidence to use in evaluations, see [50].

Evidence-based explanations are crucial for accountability and for the usefulness of the evaluation. Providing evidence encourages accountability in the evaluation process because it requires the evaluator to base the scoring on information that others can assess, too. Moreover, requiring evaluators to provide evidence also encourages accountability on the part of the evaluated companies, because it encourages them to ensure that such evidence is available. Companies can do so, for example, by documenting key processes and their outcomes.

Further, evidence-based explanations improve the usefulness of the evaluation because they contextualize and explain the reason for the score. Numbers on their own don't offer much information about the company, what they currently do, what is missing, and how they can improve. The evidence an evaluator cites helps others understand how the evaluator interprets the scoring guidelines and what a given score means to that evaluator. This can help companies understand what they are doing right and how to do better.

Applicability of the Scoring Guidelines

Inevitably, there is going to be some divergence in the scores when completed by different evaluators. This will be true for any set of guidelines, as no set of guidelines can cover all the details relevant to the wide range of contexts and circumstances evaluators may encounter. No matter how detailed the guidelines may be, evaluators will always need to exercise some judgment, deciding what satisfies the metrics and to what degree, deciding what counts as evidence, deciding which contextual factors matter most, etc.

4. SCORE AGGREGATION

After scoring the individual statements or topics, evaluators can aggregate the scoring to get a unified score. Evaluators who score individual statements can benefit from fine-grained aggregations. Our maturity model offers two modes of fine-grained aggregation: By NIST pillars and by responsibility dimensions (see Figure 2 for an illustration).

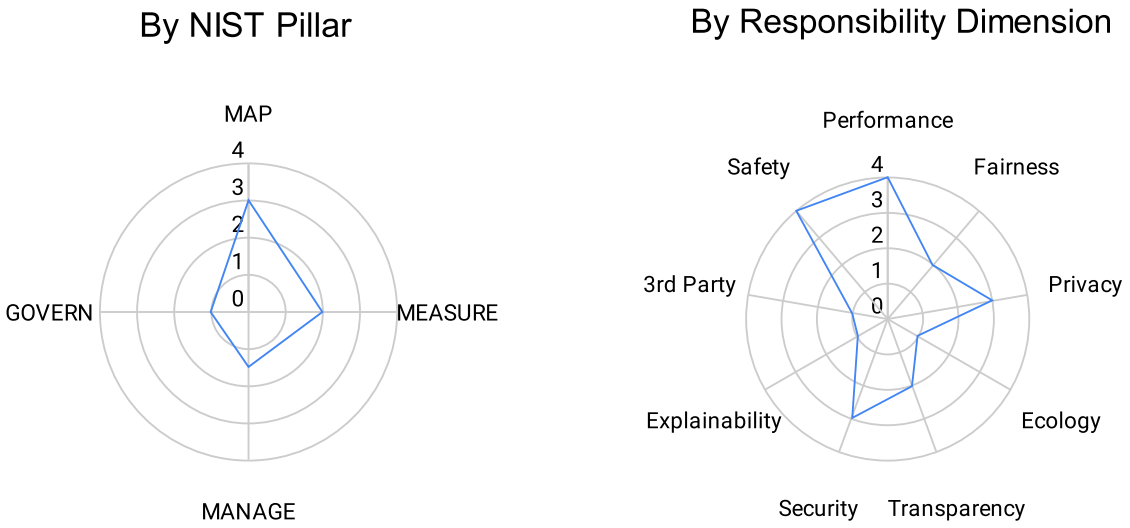


Figure 2. Illustration of aggregation modes in radio charts: To the left, aggregation by NIST Pillar. To the right, aggregation by responsibility dimension

In aggregating by NIST pillars, the output is a score for each of the NIST pillars, MAP, MEASURE, MANAGE, and GOVERN, based on the scores of the statements that belong to it. For example, the MAP score is the average of all the statements that belong to the MAP pillar (see the associations in the Appendix).

Aggregation by NIST pillar can help discover organizations' strengths and weaknesses in different kinds of activities. In particular, this mode of aggregation can expose systematic failures in organizations' approaches to AI responsibility. For example, when organizations show strength in GOVERN activities but weakness in all other pillars, they may be engaged in ethics washing. For example, they may be establishing policies that are largely not implemented. Other organizations may show strength in GOVERN and MANAGE but weakness in MAP and MEASURE. These organizations' risk management activities may be ill-informed, as the low level of MAP and MEASURE may indicate that their understanding of the risks is lacking.

Another option is to aggregate based on some or all of the dimensions of AI responsibility the RMF highlights in the subcategories — performance, fairness, privacy, ecology, transparency, security, explainability, 3rd party (e.g., IP/copyright), and safety. In this aggregation mode, the score of each dimension is an average of all the statements associated with that dimension (see the associations in the Appendix).

Aggregation by responsibility dimensions can help discover when organizations ignore certain issues. For example, some organizations boast AI responsibility based on their activity in a handful of risk areas, such as privacy and security. Focusing on each dimension can highlight the other areas in which the company is lacking.

Aggregation can help track companies' progress over time. This maturity model isn't prescriptive about the trajectory of the progress. It allows tracking progress which may take place in many different ways. For example, in large corporations, we may see a top-down progress trajectory, where the company starts with strong GOVERN activities and advances to stronger MEASURE and MANAGE activities. In smaller companies, we might see a bottom-up progress trajectory, where the company starts with strong MEASURE, MAP, and MANAGE activities and progresses to stronger GOVERN activities (see Figure 3 for an illustration).

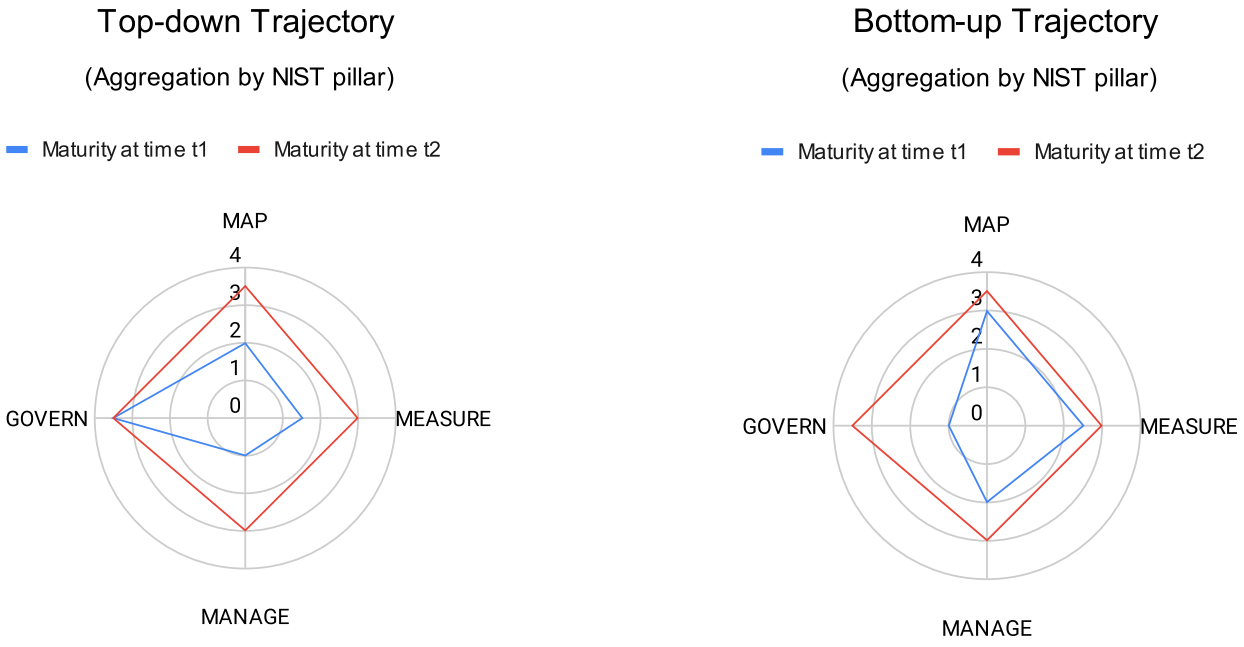


Figure 3. Illustration of maturity progress trajectories. To the left, a bottom-up trajectory. To the right, a top-town trajectory

5. CONCLUSION

This paper lays out a maturity model to evaluate the responsibility of AI governance in organizations that develop and manage AI systems. This model includes a flexible questionnaire and scoring guidelines, both based on industry standards set out by NIST. The strengths of this model include a rigorous conceptual framework that is drawn from industry standards, a focus on the mitigation of sociotechnical harm and on inclusivity, flexibility in questionnaire and aggregation options to accommodate the needs of different organizations, compatibility with multiple maturity trajectories, and the facilitation of evidence-based evaluations that flesh out subjective judgments and the reasoning to support them. All these features are intended to make this model practical and helpful in supporting organizations in improving their AI risk management and in supporting the field in enhancing the overall levels of AI ethics implementation. For more about the maturity model and its implementation, see [49] and [50].

6. ACKNOWLEDGMENTS

Ravit Dotan was supported by a grant from the Notre Dame-IBM Tech Ethics Lab.

We are thankful to our collaborators on applied aspects of this ongoing project for extensive feedback on this paper: Benny Esparra, and Ric McLaughlin. We also thank the following individuals and organizations for their help and feedback: (in alphabetical order) Anita Dorett, Jesse Dunietz, Diana Glassman, Saurabh Gupta (KOKO Labs), Javier Lempert (Light-It), Adam Mallat (Light-It), Navishka Pandit, Larisa Ruoff, and Reva Schwartz.

APPENDICES

The Questionnaire

In this appendix, we present the maturity model's questionnaire. As discussed in the body of the paper, the questionnaire is divided by stages of the development life-cycle. Altogether, the questionnaire contains 9 topic statements with more detailed individual statements under each topic statement. There are a total of 66 detailed statements. Evaluators can choose to evaluate topic statements only or all statements. They are asked to provide both a score and an explanation to support each score.

Figure 1 provides a visual illustration of the questionnaire's structure, and Tables A.1-A.3 contain the list of all the statements (topic and detailed). Each statement is associated with responsibility dimension (performance, fairness, privacy, ecology, transparency, security, explainability, 3rd party, and safety), specific NIST subcategories, and a NIST Pillar (MAP, GOV, MEA, MAN). Some of the statements reflect multiple NIST subcategories, sometimes from different pillars. Each statement was associated with one of the pillars of its subcategories, based on how well the statement matches the spirit of each pillar:

- **MAP** - Learning about AI risks and opportunities
- **MEASURE** - Measuring risks and impacts
- **MANAGE** - Implementing practices to mitigate risks and maximize benefits
- **GOVERN** - Systematizing and organizing activities across the organization.

The questionnaire has changed in some ways relative to [49]. In particular, there is less of an emphasis on documentation based on experience using the questionnaire in practice [50]. The exact details of the questionnaire may vary over time or with specific use cases. This is not presented as a single definitive set of questions.

Table A. 1 Topics Relevant for Systems in the Planning Stage and Beyond (All Systems)			
NIST Pillar	NIST Subcategories	Responsibility Dimension	**1. Mapping impacts** We clearly define what the AI is supposed to do and its impacts, including scope, goals, methods, and the negative and positive potential impacts of these activities.
MAP	MAP 1.3, MAP 2.1, MAP 3.3		**1.1 Goals** We define the **goals, scope, and methods** of this AI system.
MAP	MAP 1.1, MAP 3.1, MAP 5.1, GOV 4.2		**1.2 Positive Impacts** We identify the **benefits and potential positive impacts** of this AI system, including the likelihood and magnitude.
MAP	MAP 1.4, MAP 3.1		**1.3 Business Value** We identify the **business value** of this AI system
MAP	MAP 5.1, GOV 4.2		**1.4 Negative impacts** We identify the possible **negative impacts** of this AI system, including the likelihood and magnitude.
MAP	MAP 3.2		**1.5 Costs of malfunction** We identify the potential **costs** of malfunctions of this AI system, including non-monetary costs such as decreased trustworthiness
MAP	MAP 5.2	Unexpected	**1. 6 Unexpected Impacts** We implement processes to integrate input about **unexpected impacts**
MAP	MAP 2.3, MAP 4.1		**1.7 Methods and tools** We document the **methods and tools** we use for mapping impacts
MAP	MAP 1.2, GOV 3.1, GOV 5.1, GOV 5.2	Input diversity	**1.8 Diverse input** **Diverse stakeholders inform** the mapping process, including diverse skills and demographic backgrounds
			2. Identifying requirements We identify the requirements the AI must meet, including compliance, certifications, and human oversight needs.

GOV	GOV 3.2, MAP 3.5	Oversight	**2.1 Human oversight** We identify the **human oversight** processes the system needs
MAP	MAP 1.6, MAP 3.4		**2.2 Standards** We identify the **technical standards and certifications** the system will need to satisfy
GOV	GOV 1.1		**2.3 Legal** We identify AI **legal requirements** that apply to this AI system
			3. AI ethics mindset and culture We facilitate a mindset of responsibility, for example, by providing AI ethics training to relevant personnel, clearly defining relevant roles, establishing policies, and implementing practices for critical thinking.
GOV	GOV 1.2, GOV 1.4		**3.1 Policies** We write policies and guidelines about AI ethics
GOV	GOV 2.1		**3.2 Roles** We document **roles, responsibilities, and lines of communication** related to AI risk management
GOV	GOV 2.2		**3.3 Training** We provide **training** about AI ethics to relevant personnel
GOV	GOV 4.1		**3.4 Critical Thinking** We implement practices to foster **critical thinking** about AI risks
GOV	GOV 2.3		**3.5 Leadership** Executive **leadership takes responsibility** for decisions related to AI risks

Table A. 2 Topics Relevant for Systems in the Data Collection and Model Building Stage and Beyond

NIST Pillar	NIST Subcategories	Responsibility Dimension	**4. Measuring risk** We measure potential negative impacts.
			4.1 Strategy We make and periodically re-evaluate our strategy for measuring the impacts of this AI system. It includes choosing which impacts we measure. It also includes how we will approach monitoring unexpected impacts and impacts that can't be captured with existing metrics.
MEA	MEA 1.2, MEA 2.1, MEA 3.1, MEA 3.2, MAP 2.3		**4.2 Methods** We have a clear set of methods and tools we use to measure the impacts of this AI system. It includes which metrics and datasets we use.
MEA	MEA 1.2, MEA 2.13		**4.3 Effectiveness** We evaluate the effectiveness of our measurement processes
MEA	MEA 2.3	Performance	**4.4 Performance** We regularly evaluate the performance of this AI system in conditions similar to deployment
MEA	MEA 2.11	Fairness	**4.5 Bias & fairness** We regularly evaluate bias and fairness issues related to this AI system
MEA	MEA 2.10	Privacy	**4.6 Privacy** We regularly evaluate privacy issues related to this AI system
MEA	MEA 2.12	Ecology	**4.7 Environment** We regularly evaluate environmental impacts related to this AI system
MEA	MEA 2.8	Transparency	**4.8 Transparency & Accountability** We regularly evaluate transparency and accountability issues related to this AI system
MEA	MEA 2.7	Security	**4.9 Security** We regularly evaluate security and resilience issues related to this AI system

MEA	MEA 2.9	Explainability	**4.10 Explainability** We regularly evaluate explainability issues related to this AI system
MEA	MEA 1.1, GOV 6.1	3rd party	**4.11 3rd Party** We regularly evaluate third-party issues, such as IP infringement, related to this AI system
MEA	MEA 1.1, MAP 3.5, GOV 3.2	Oversight	**4.12 Human oversight** We regularly evaluate human oversight issues related to this AI system
MEA	MEA 2.6	Safety	**4.13 Safety** We regularly evaluate safety issues related to this AI system
MEA	MEA 1.1	Other	**4.14 Other** We regularly evaluate other impacts related to this AI system
MEA	MEA 2.2, MEA 2.6		**4.15 Human subjects** If evaluations use human subjects, they are representative and meet appropriate requirements
MEA	MEA 1.3, MEA 3.3, MEA 4.1, MEA 4.2, MEA 4.3, GOV 5.2, GOV 3.1	Input diversity	**4.16 Diverse input** Consultations with diverse domain experts and end users inform measurement approaches, results, and progress.
			5. Transparency We document information about the system, including explaining how it works, limitations, and risk controls.
GOV	GOV 1.4, MAP 2.2	Transparency	**5.1 Limitations & Oversight** We document information about the system's limitations and options for human oversight related to this AI system. The documentation is good enough to assist those who need to make decisions based on the system's outputs.
GOV	GOV 1.4, MAP 4.2	Transparency	**5.2 Risk controls** We document the system risk controls, including in third-party components

GOV	GOV 1.4, MEA 2.9	Explainability	**5.3 Model explanation** We explain the model to ensure responsible use We inventory information about this AI system in a repository of our AI systems
GOV	GOV 1.6	Transparency	
			6. Management plan We plan how to respond to risks, including setting priorities and documenting residual risks.
GOV	GOV 1.3, GOV 1.4, MAP 1.5, MAN 1.3, MAN 2.1		**6.1 Plan** We plan how we will respond to the risks caused by this AI system. The response options include defining the organization's risk tolerance level and deciding when risks should be mitigated, avoided, or accepted.
GOV	GOV 1.3, MAN 1.2		**6.2 Prioritization** We prioritize the responses to the risks of this AI system based on impact, likelihood, available resources or methods, and the organization's risk tolerance.
MAP	MAN 1.4		**6.3 Residual risk** We identify the residual risks of this AI system (the risks that we do not mitigate), including risks to buyers and users of the system.
GOV	GOV 1.4, MAN 1.2, MAN 2.3	Unexpected	**6.4 Unexpected risks** We have a plan for addressing unexpected risks related to this AI system as they come up
			7. Risk mitigation We act to minimize risks, including addressing your prioritized risks and tracking incidents.
MAN	MAN 1.1, GOV 1.5	Performance	**7.1 Meets objectives** We proactively evaluate whether this system meets its stated objectives and whether its development or deployment should proceed
MAN	MAN 1.3, MAN 4.2, MEA 2.11	Fairness	**7.2 Bias & fairness** We ensure this AI's bias and fairness performance stays meets our standards
MAN	MAN 1.3, MEA 2.10, MAN 4.2	Privacy	**7.3 Privacy** We ensure this AI's privacy performance meets our standards

MAN	MAN 1.3, MEA 2.12, MAN 4.2	Ecology	**7.4 Environment** We ensure this AI's environmental performance meets our standards
MAN	MAN 1.3, MEA 2.8, MAN 4.2	Transparency	**7.5 Transparency & Accountability** We ensure this AI's transparency and accountability meets our standards
MAN	MAN 1.3, MEA 2.7, MAN 4.2	Security	**7.6 Security** We ensure this AI's security and resilience meets our standards
MAN	MAN 1.3, MEA 2.9, MAN 4.2	Explainability	**7.7 Explainability** We ensure this AI's explainability performance meets our standards
MAN	MAN 3.1, GOV 6.1, MAN 1.3	3rd party	**7.8 3rd party** We ensure this AI's third-party impacts, such as IP infringement, meet our standards
MAN	GOV 3.2, MAP 3.5, MAN 1.3	Oversight	**7.9 Human oversight** We implement processes for human oversight related to this AI system
MAN	MAN 4.1	Oversight	**7.10 Appeal** We implement processes for appeal related to this AI system
MAN	MAN 2.4, GOV 1.7		**7.11 End of life** We maintain end-of-life mechanisms to supersede, disengage, or deactivate this AI system if its performance or outcomes are inconsistent with the intended use.
MAN	MAN 1.3, MEA 2.6, MAN 4.2	Safety	**7.12 Safety** We ensure this AI system is safe
MAN	MAN 1.3, MAN 4.2	Other	**7.13 Other risks** We address all other risks prioritized in our plans related to this system by conducting measurable activities
MAN	MAN 2.3	Unexpected	**7.14 Unexpected risks** We address unexpected risks related to this system by conducting measurable activities

MAN	MAN 4.3, GOV 4.3		**7.15 Errors and incidents** We track and respond to errors and incidents related to this system by conducting measurable activities
GOV	MEA 1.3, MEA 3.3, MEA 4.1, MEA 4.2, MEA 4.3, GOV 5.2, GOV 3.1	Input diversity	**7.16 Input diversity** Consultations with diverse domain experts and end users inform risk management activities

Table A. 3 Topics Relevant for Systems in the Deployment Stage and Beyond

NIST Pillar	NIST Subcategories	Responsibility Dimension	**8. Pre-deployment checks** We only release features that meet our AI ethics standards.
MAN	MAN 1.1, MEA 2.5		**8.1 Valid and reliable** We demonstrate that this system and its features are **valid, reliable, and meet our standards**. We document the conditions under which it falls short.
			9. Monitoring We monitor and resolve issues as they arise.
GOV	MAN 4.1, GOV 1.3		**9.1 Monitoring Plan** We **plan how to monitor** risks related to this system post-deployment
MEA	MEA 2.4, MEA 2.6	Performance	**9.2 Functionality** We monitor this system's **functionality and behavior** post-deployment
MAN	MAN 2.2		**9.3 Sustain Value** We apply mechanisms to **sustain the value of this AI system** post-deployment
MAN	MAN 4.1, GOV 5.2	Input diversity	**9.4 User Input** We capture and evaluate **input from users** about this system post-deployment
MAN	MAN 4.1, MEA 2.6	Oversight	**9.5 Appeal and override** We monitor **appeal and override** processes related to this system post-deployment
MAN	MAN 4.1, GOV 4.3, MEA 2.6		**9.6 Incidents** We monitor **incidents** related to this system **and responses** to them post-deployment
MAN	GOV 6.2, MEA 2.6, MAN 3.1, MAN 3.2	3rd party	**9.7 3rd Party** We monitor incidents related to **third-party** components, such as pre-trained models or data, and respond to them, especially when these components are high risk

MAN	MAN 4.1, MEA 2.6		**9.8 Other** We implement **all other** components of our post-deployment monitoring plan for this system
MAN	MAN 2.4, MAN 4.1, MEA 2.6		**9.9 End of life** We monitor issues that would trigger our **end-of-life mechanisms** for this system, and we take the system offline if issues come up

REFERENCES

[1] C. Drew, "Design for data ethics: Using service design approaches to operationalize ethical principles on four projects," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 376, no. 2128, p. 20170353, Sep. 2018, doi: 10.1098/rsta.2017.0353.

[2] J. Morley, L. Floridi, L. Kinsey, and A. Elhalal, "From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices," *Science and Engineering Ethics*, vol. 26, no. 4, pp. 2141–2168, Aug. 2020, doi: 10.1007/s11948-019-00165-5.

[3] J. Morley, L. Kinsey, A. Elhalal, F. Garcia, M. Ziosi, and L. Floridi, "Operationalising AI ethics: Barriers, enablers and next steps," *AI & SOCIETY*, vol. 38, no. 1, pp. 411–423, Feb. 2023, doi: 10.1007/s00146-021-01308-8.

[4] J. Ayling and A. Chapman, "Putting AI ethics to work: Are the tools fit for purpose?" *AI and Ethics*, vol. 2, no. 3, pp. 405–429, Aug. 2022, doi: 10.1007/s43681-021-00084-x.

[5] L. Zhu, X. Xu, Q. Lu, G. Governatori, and J. Whittle, "AI and Ethics – Operationalising Responsible AI." arXiv, May 2021. doi: 10.48550/arXiv.2105.08867.

[6] L. Munn, "The uselessness of AI ethics," *AI and Ethics*, vol. 3, no. 3, pp. 869–877, Aug. 2023, doi: 10.1007/s43681-022-00209-w.

[7] S. Holland, A. Hosny, S. Newman, J. Joseph, and K. Chmielinski, "The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards." arXiv, May 2018. doi: 10.48550/arXiv.1805.03677.

[8] T. Gebru *et al.*, "Datasheets for datasets," *Communications of the ACM*, vol. 64, no. 12, pp. 86–92, Dec. 2021, doi: 10.1145/3458723.

[9] M. Mitchell *et al.*, "Model Cards for Model Reporting," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Jan. 2019, pp. 220–229. doi: 10.1145/3287560.3287596.

[10] NIST, "AI Risk Management Framework: AI RMF (1.0)," National Institute of Standards; Technology, Gaithersburg, MD, NIST AI 100-1, 2023. doi: 10.6028/NIST.AI.100-1.

[11] IBM, "IBM Global AI Adoption Index 2022 IBM." Accessed: Jan. 18, 2024. [Online]. Available: https://www.ibm.com/watson/resources/ai-adoption

[12] McKinsey, "The state of AI in 2022—and a half decade in review McKinsey." Accessed: Jan. 18, 2024. [Online]. Available: https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2022-and-a-half-decade-in-review#/download//~/media/mckinsey/business%20functions/quantumblack/our%20insights/the%20state%20of%20ai%20in%202022%20and%20a%20half%20decade%20in%20review/the-state-of-ai-in-2022-and-a-half-decade-in-review.pdf?cid=soc-web

[13] R. Dotan, G. Rosenthal, T. Buckley, J. Scarpino, L. Patterson, and T. Bristow, "Evaluating AI Governance: Insights from Public Disclosures," 2024. Available: https://www.ravitdotan.com/_files/ugd/f83391_b853450bcc274e9ba9454d618ee41a94.pdf

[14] J. Pöppelbuß and M. Röglinger, "WHAT MAKES A USEFUL MATURITY MODEL? A FRAMEWORK OF GENERAL DESIGN PRINCIPLES FOR MATURITY MODELS AND ITS DEMONSTRATION IN BUSINESS PROCESS MANAGEMENT," *ECIS 2011 Proceedings*, Oct. 2011, Available: https://aisel.aisnet.org/ecis2011/28

[15] V. Vakkuri *et al.*, "Time for AI (Ethics) Maturity Model Is Now." arXiv, Jan. 2021. Accessed: Jan. 18, 2024. [Online]. Available: http://arxiv.org/abs/2101.12701

[16] R. Bommasani, "Evaluation for Change," 2022, doi: 10.48550/ARXIV.2212.11670.

[17] US Department of Energy, "Cybersecurity Capability Maturity Model (C2M2)," US Department of Energy, 2.1, Jun. 2022. Accessed: Jan. 18, 2024. [Online]. Available: https://www.energy.gov/ceser/cybersecurity-capability-maturity-model-c2m2

[18] R. Wendler, "The maturity of maturity model research: A systematic mapping study," *Information and Software Technology*, vol. 54, no. 12, pp. 1317–1339, Dec. 2012, doi: 10.1016/j.infsof.2012.07.007.

[19] J. Piaget, "Cognitive development in children: Piaget development and learning," *Journal of Research in Science Teaching*, vol. 2, no. 3, pp. 176–186, Sep. 1964, doi: 10.1002/tea.3660020306.

[20] S. S. Kuznets, *Economic growth and structure: Selected essays*. New York, N.Y: Norton, 1965.

[21] P. B. Crosby, *Quality is free: The art of making quality certain*. New York: McGraw-Hill, 1979.

[22] D. Proenca, J. Estevens, R. Vieira, and J. Borbinha, "Risk Management: A Maturity Model Based on ISO 31000," in *2017 IEEE 19th Conference on Business Informatics (CBI)*, Thessaloniki, Greece: IEEE, Jul. 2017, pp. 99–108. doi: 10.1109/CBI.2017.40.

[23] "CMMI Institute - Home." Accessed: Jan. 18, 2024. [Online]. Available: https://cmmiinstitute.com/

[24] NIST, NCCoE, DOE, and CESER, "Cybersecurity Capability Maturity Model to NIST Cybersecurity Framework Mapping NCCoE." Mar. 2023. Accessed: Jan. 18, 2024. [Online]. Available: https://www.nccoe.nist.gov/news-insights/cybersecurity-capability-maturity-model-nist-cybersecurity-framework-mapping

[25] National Institute of Standards and Technology, "NIST PRIVACY FRAMEWORK:: A TOOL FOR IMPROVING PRIVACY THROUGH ENTERPRISE RISK MANAGEMENT, VERSION 1.0," National Institute of Standards; Technology, Gaithersburg, MD, NIST CSWP 01162020, Jan. 2020. doi: 10.6028/NIST.CSWP.01162020.

[26] CCMC, "CMMC PROGRAM PROPOSED RULE PUBLISHED - PUBLIC COMMENT PERIOD BEGINS," US Department of Defense, 2021. Accessed: Jan. 18, 2024. [Online]. Available: https://dodcio.defense.gov/CMMC/about/

[27] Baxter, K, "AI Ethics Maturity Model," Salesforce, 2021. Accessed: Jan. 18, 2024. [Online]. Available: https://www.salesforceairesearch.com/static/ethics/EthicalAIMaturityModel.pdf

[28] M. Vorvoreanu, A. Heger, S. Passi, S. Dhanorkar, Z. Kahn, and R. Wang, "Responsible AI Maturity

Model," Microsoft, MSR-TR-2023-26, May 2023. Available: https://www.microsoft.com/en-us/research/publication/responsible-ai-maturity-model/

[29] Open Data Institute, "Data Ethics Maturity Model: Benchmarking your approach to data ethics," Mar. 2022. Accessed: Jan. 18, 2024. [Online]. Available: https://theodi.org/insights/tools/data-ethics-maturity-model-benchmarking-your-approach-to-data-ethics/

[30] Ethical Intelligence, BCV, and EAIGG, "ETHICS MATURITY CONTINUUM," 2022. Accessed: Jan. 18, 2024. [Online]. Available: https://static1.squarespace.com/static/5f6dbf464a8eec79c3d177c0/t/61e8821d53b74041072d556d/1642627614838/Ethics+Maturity+Continuum+Report.pdf

[31] J. Krijger, T. Thuis, M. de Ruiter, E. Ligthart, and I. Broekman, "The AI ethics maturity model: A holistic approach to advancing ethical data science in organizations," *AI and Ethics*, vol. 3, no. 2, pp. 355–367, May 2023, doi: 10.1007/s43681-022-00228-7.

[32] IBM, "AI maturity framework for enterprise applications," IBM, Oct. 2021. Accessed: Jan. 18, 2024. [Online]. Available: https://www.ibm.com/watson/supply-chain/resources/ai-maturity

[33] MITRE, "The MITRE AI Maturity Model and Organizational Assessment Tool Guide: A Path to Successful AI Adoption," MITRE, 2023.

[34] PwC, "Responsible AI - Maturing from theory to practice," PwC, 2021.

[35] C. Covello and K. Iatridis, "On the challenges and drivers of implementing responsible innovation in foodpreneurial SMEs," in *Assessment of responsible innovation*, 1st ed., E. Yaghmaei and I. van de Poel, Eds., Routledge, 2020. doi: https://doi.org/10.4324/9780429298998.

[36] H. E. J. Bos-Brouwers, "Corporate sustainability and innovation in SMEs: Evidence of themes and activities in practice," *Business Strategy and the Environment*, vol. 19, pp. 417–435, Jun. 2009, doi: https://doi.org/10.1002/bse.652.

[37] C. Oprysko, "OpenAI registers to lobby," *Politico*, Nov. 2023, Accessed: Jan. 22, 2024. [Online]. Available: https://www.politico.com/newsletters/politico-influence/2023/11/17/openai-registers-to-lobby-00127874

[38] The White House, "[Executive Order 14110 of October 30, 2023] Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence." Oct. 2023. Accessed: Jan. 18, 2024. [Online]. Available: https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence

[39] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi, "Fairness and Abstraction in Sociotechnical Systems," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Atlanta GA USA: ACM, Jan. 2019, pp. 59–68. doi: 10.1145/3287560.3287598.

[40] L. Weidinger *et al.*, "Sociotechnical Safety Evaluation of Generative AI Systems." arXiv, Oct. 2023. Accessed: Nov. 02, 2023. [Online]. Available: http://arxiv.org/abs/2310.11986

[41] S. Lazar and A. Nelson, "AI safety on whose terms?" *Science*, vol. 381, no. 6654, pp. 138–138, Jul. 2023, doi: 10.1126/science.adi8982.

[42] C. A. E. Goodhart, "Problems of Monetary Management: The UK Experience," in *Monetary Theory and Practice*, London: Macmillan Education UK, 1984, pp. 91–121. doi: 10.1007/978-1-349-17295-5_4.

[43] C. Hennessy and C. A. E. Goodhart, "Goodhart's Law and Machine Learning," *SSRN Electronic Journal*, 2020, doi: 10.2139/ssrn.3639508.

[44] A. Chrystal and P. Mizen, "Goodhart's Law: Its origins, meaning and implications for monetary policy," in *Central Banking, Monetary Theory and Practice*, Edward Elgar Publishing, 2003, p. 2329. doi: 10.4337/9781781950777.00022.

[45] M. Strathern, "'Improving ratings': Audit in the British University system," *European Review*, vol. 5, no. 3, pp. 305–321, Jul. 1997, doi: 10.1002/(SICI)1234-981X(199707)5:3<305::AID-EURO184>3.0.CO;2-4.

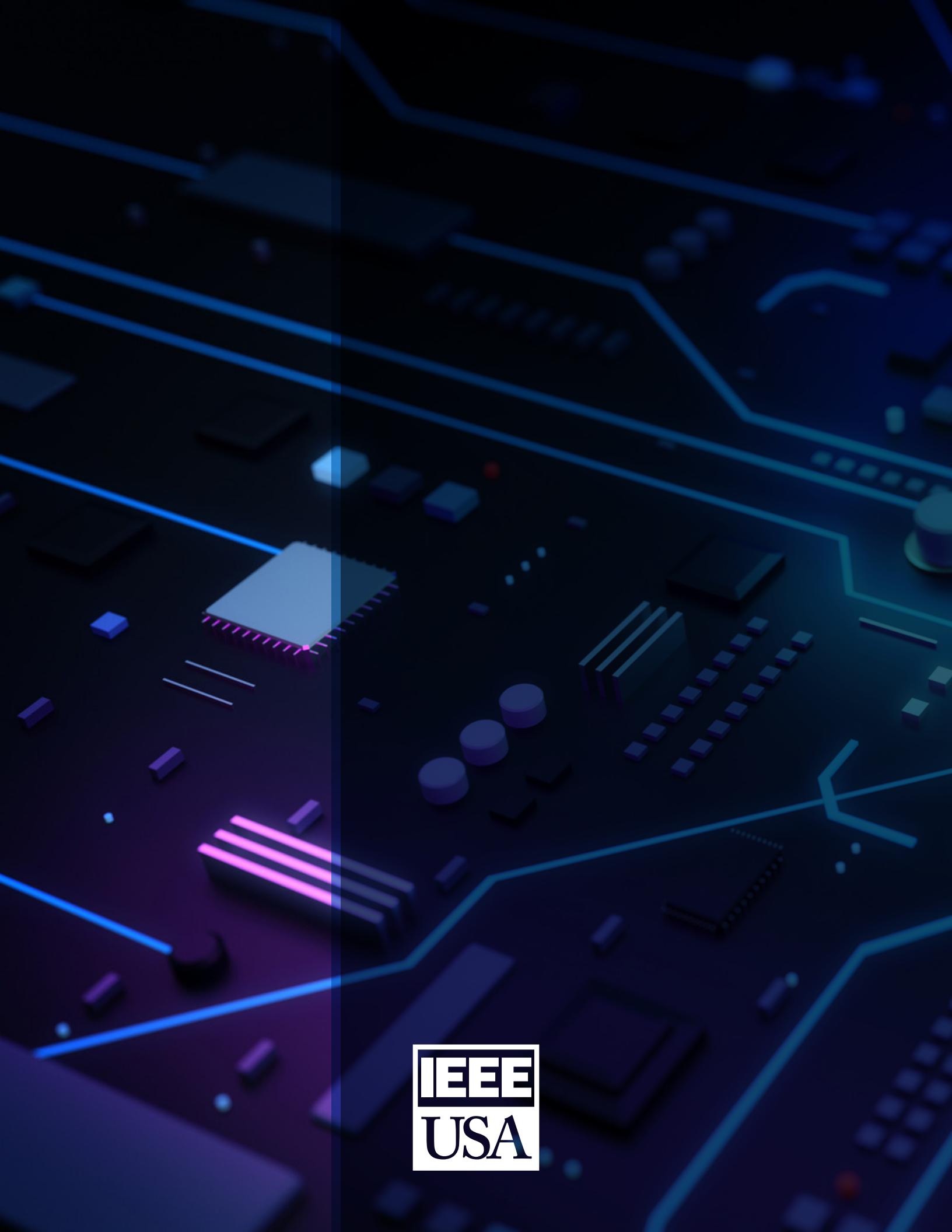
[46] OpenAI, "Our approach to AI safety," *OpenAI*. Apr. 2023. Accessed: Jan. 18, 2024. [Online]. Available: https://openai.com/blog/our-approach-to-ai-safety

[47] Burstein, Jill, "Duolingo English Test Responsible AI Standards," Duolingo, 2023. Accessed: Jan. 18, 2024. [Online]. Available: https://duolingo-papers.s3.amazonaws.com/other/DET+Responsible+AI+033123.pdf

[48] OpenAI, "Preparedness," *OpenAI*. Accessed: Jan. 18, 2024. [Online]. Available: https://openai.com/safety/preparedness

[49] R. Dotan, B. Blili-Hamelin, R. Madhavan, J. Matthews, J. Scarpino. "Evolving AI Risk Management: A Maturity Model based on the NIST AI Risk Management Framework". Accessed: April 27, 2024. [Online]. Available: https://arxiv.org/abs/2401.15229

[50] R. Dotan, B. Blili-Hamelin, R. Madhavan, J. Matthews, J. Scarpino, C. Anderson, R. McLaughlin, B. Esparra. "Responsible AI Governance Maturity Model: 2024 Hackathon Report". *All Tech is Human*. Accessed: April 27, 2024. [Online]. Available: https://www.techbetter.ai/rai-maturity-model


IEEE
USA