all tech is
human

Responsible AI Governance Maturity Model

2024 Hackathon Report

Authors: Ravit Dotan, Borhane Blili-Hamelin, Ravi Madhavan, Jeanna Matthews, Joshua Scarpino, Carol Anderson, Benny Esparra, and Ric Mclaughlin

Foreword

The Responsible AI Governance Maturity Model is a framework for evaluating the social responsibility of AI governance in organizations that develop or use AI systems, based on the NIST AI Risk Management Framework. When used internally, the model is intended to help organizations diagnose where they stand and improve. When used externally, e.g., by investors, buyers, or consumers, the model is intended to inform decision-making about the company and its products.

The maturity model was developed in a research paper (Dotan et al. 2024) by Ravit Dotan, Borhane Blili-Hamelin, Ravi Madhavan, Jeanna Matthews, and Joshua Scarpino. Together with Carol Anderson, Ric Mclaughlin, and Benny Esparra we partnered with All Tech is Human and created a hackathon for engaging with the maturity model.

The goal of the hackathon was to bring diverse perspectives into the development of the model, and, at the same time, provide AI ethics education and networking opportunities. During the hackathon, participants practiced AI ethics skills and used the maturity model to evaluate either their own company using internal information or any other company using public information. The hackathon was a great success, with 250 registrants, 180 active participants, and 54 submitted evaluations. The hackathon team used participants' work and feedback to improve the model.

In this report, we describe the maturity model, the hackathon, and learnings about the maturity model arising from the participants' work. Thank you to all the participants and to the hackathon funder, The Notre Dame-IBM Tech Ethics Lab!

RAVIT DOTAN

Hackathon Lead
TechBetter, CEO

REBEKAH TWEED

Executive Director
All Tech Is Human

Table of Contents

About
The Responsible AI Governance Maturity Model

About
The Responsible AI Governance Maturity Model

The gap:
Why organizations need a maturity model for responsible AI governance

Organizations are struggling with AI governance, to the point that it often even hinders adoption. For example, in a recent BCG survey, 52% of executives said that they actively discourage the use of generative AI. The lack of a responsible AI strategy was the second most common reason for the discouragement, and the most common concerns included data breaches, unpredictable outcomes, and making wrong or biased decisions.

Maturity models are tools that help organizations improve their capabilities. They first appeared in software development (e.g. the SEI capability model) and have now spread to many other areas such as knowledge management, cybersecurity, and innovation. From a managerial perspective, maturity models chart an evolutionary path toward greater capability, with associated tools for assessing where the enterprise stands and providing guidance on what changes are necessary for continuous improvement.

The Responsible AI Governance maturity model helps organizations plan and improve their capabilities in AI risk management. It includes a questionnaire and scoring guidelines. We describe them briefly below. For the full version, see the appendix and the research paper in which the model was developed: Dotan et al. (2024).

The Foundation: The NIST AI RMF

The maturity model is based on the NIST AI Risk Management Framework (AI RMF), one of the most well-respected AI governance frameworks in the world.

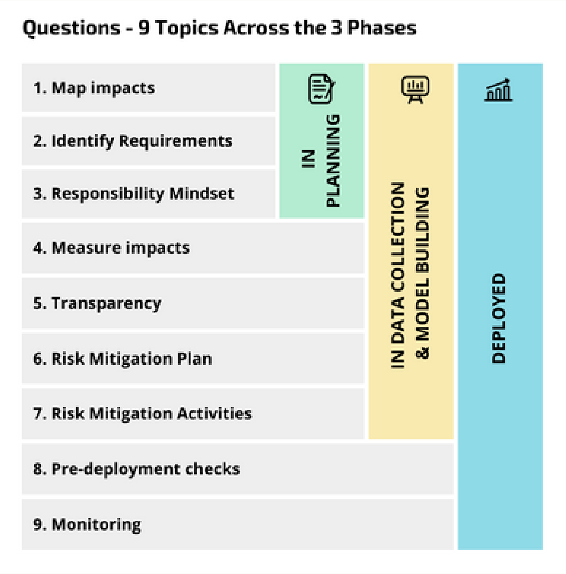
The NIST AI RMF recommends best practices for AI risk management, listing many activities for developing and deploying AI in socially responsible ways. While the NIST AI RMF is comprehensive, it does not provide guidance on how organizations can measure their degree of alignment or evaluate progress. **The maturity model bridges the gap to implementation by offering a way to assess and track the evolution of an organization's AI governance.**

The Questionnaire

The questionnaire is composed of a list of statements that is divided into nine topics.

The topics are organized into stages of the development life-cycle (based on the NIST categorization). The evaluator only scores the statements suitable for the life-cycle stage of the AI system or the company.

For example, during model-building phases, evaluators only use topics 1-7.

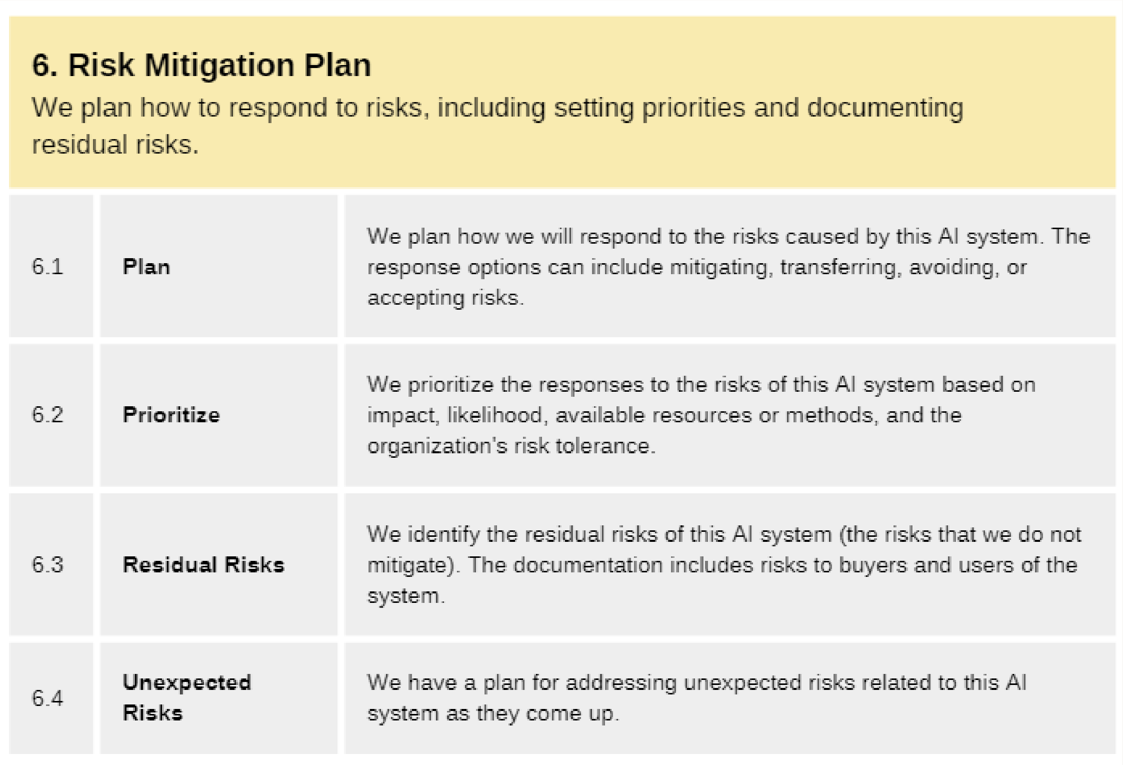


Questions - 9 Topics Across the 3 Phases

1. Map impacts
2. Identify Requirements
3. Responsibility Mindset
4. Measure impacts
5. Transparency
6. Risk Mitigation Plan
7. Risk Mitigation Activities
8. Pre-deployment checks
9. Monitoring

Each topic contains a list of statements. For example, the topic "4. Measure impacts" includes statements such as "We evaluate bias and fairness issues caused by our AI systems." In isolation, each statement covers one or more of the recommended NIST AI RMF activities. Jointly, they cover all the recommendations. Further, the statements center on concrete and verifiable actions companies may perform, avoiding general and abstract statements such as "Our AI systems are fair."

You can find the full questionnaire in the appendix.



6. Risk Mitigation Plan

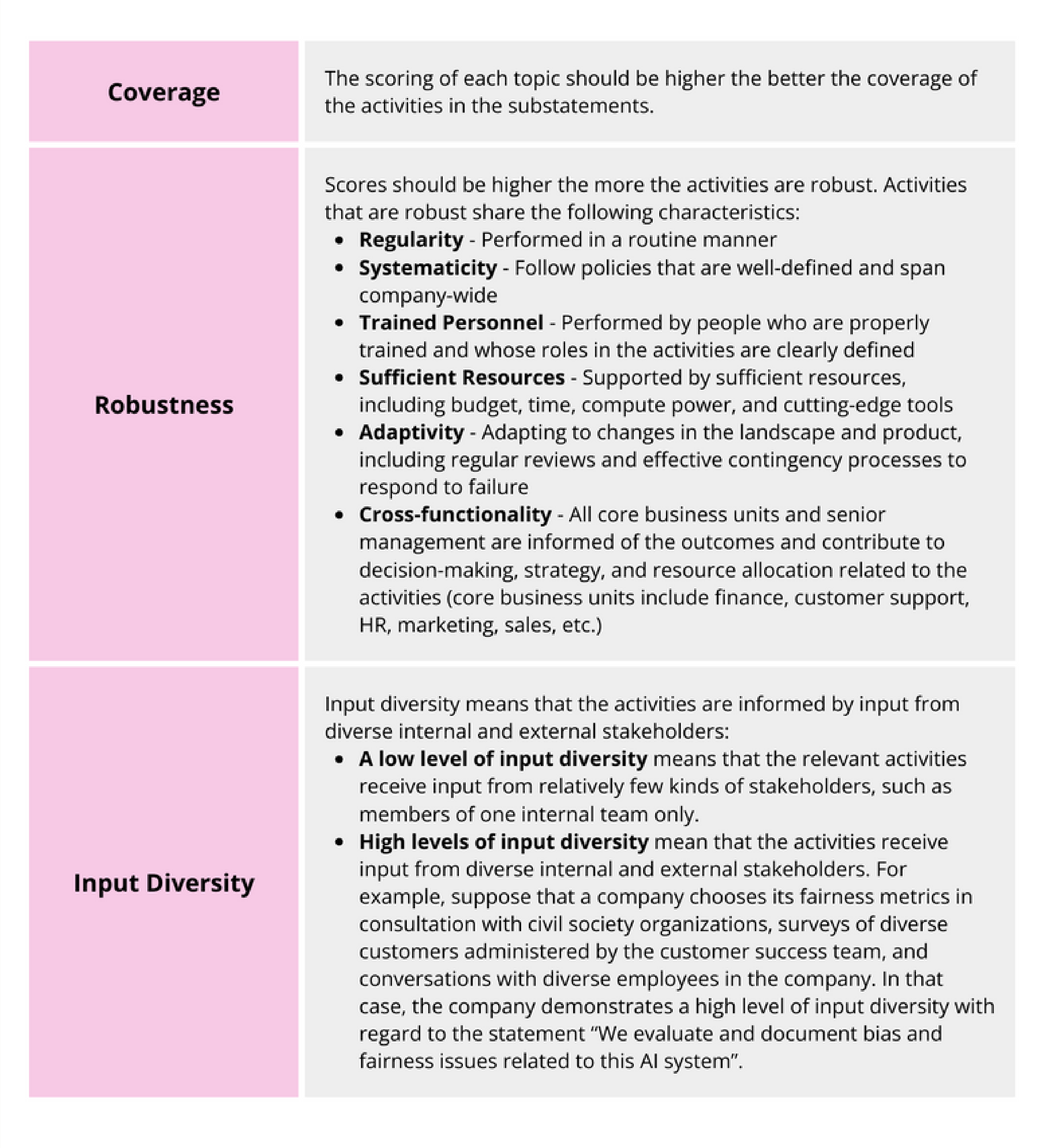
We plan how to respond to risks, including setting priorities and documenting residual risks.

6.1	**Plan**	We plan how we will respond to the risks caused by this AI system. The response options can include mitigating, transferring, avoiding, or accepting risks.
6.2	**Prioritize**	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.
6.3	**Residual Risks**	We identify the residual risks of this AI system (the risks that we do not mitigate). The documentation includes risks to buyers and users of the system.
6.4	**Unexpected Risks**	We have a plan for addressing unexpected risks related to this AI system as they come up.

Scoring Guidelines

The scoring guidelines build upon ideals from the NIST AI RMF as well as another NIST resource: implementation tiers (e.g., see NIST's Privacy Framework and Cybersecurity Framework).

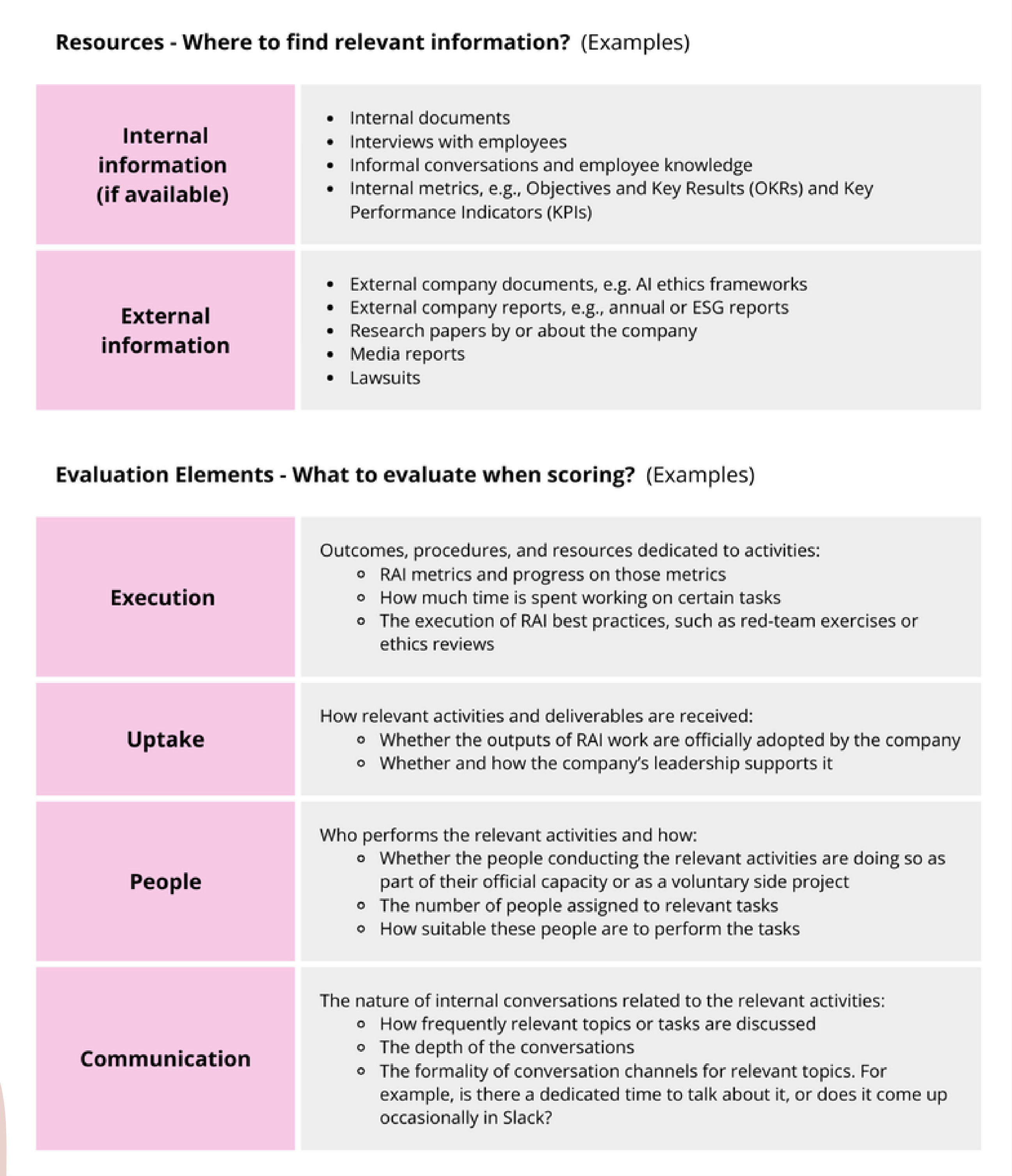
From these resources, we have extracted three concepts that should guide the scoring: **Coverage, Robustness, and Input Diversity**. The evaluator ranks each topic on these three metrics, and the overall score is calculated automatically based on this ranking. You can find the full scoring guidelines in the appendix.



Coverage	The scoring of each topic should be higher the better the coverage of the activities in the substatements.
Robustness	Scores should be higher the more the activities are robust. Activities that are robust share the following characteristics: • **Regularity** - Performed in a routine manner • **Systematicity** - Follow policies that are well-defined and span company-wide • **Trained Personnel** - Performed by people who are properly trained and whose roles in the activities are clearly defined • **Sufficient Resources** - Supported by sufficient resources, including budget, time, compute power, and cutting-edge tools • **Adaptivity** - Adapting to changes in the landscape and product, including regular reviews and effective contingency processes to respond to failure • **Cross-functionality** - All core business units and senior management 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 the activities are informed by 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, such as members of one internal team only. • **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".

Explaining Scores

Scores must be accompanied by explanations. The explanation should refer to information about what the organization does or doesn't do and any relevant contextual facts. Below are examples from the hackathon for information to use when scoring, evaluation elements, and where that information may be found, i.e., resources.



Resources - Where to find relevant information? (Examples)

Internal information (if available)	• Internal documents • Interviews with employees • Informal conversations and employee knowledge • Internal metrics, e.g., Objectives and Key Results (OKRs) and Key Performance Indicators (KPIs)
External information	• External company documents, e.g. AI ethics frameworks • External company reports, e.g., annual or ESG reports • Research papers by or about the company • Media reports • Lawsuits

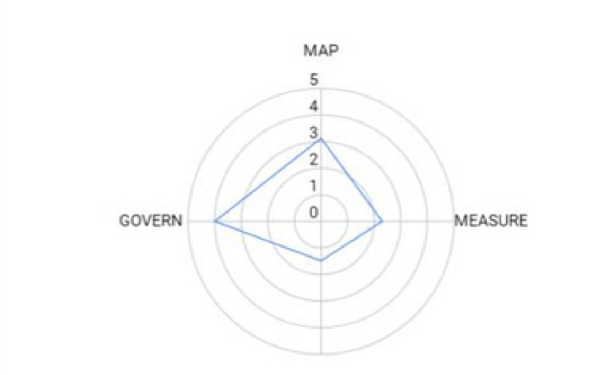
Evaluation Elements - What to evaluate when scoring? (Examples)

Execution	Outcomes, procedures, and resources dedicated to activities: ○ RAI metrics and progress on those metrics ○ How much time is spent working on certain tasks ○ The execution of RAI best practices, such as red-team exercises or ethics reviews
Uptake	How relevant activities and deliverables are received: ○ Whether the outputs of RAI work are officially adopted by the company ○ Whether and how the company's leadership supports it
People	Who performs the relevant activities and how: ○ Whether the people conducting the relevant activities are doing so as part of their official capacity or as a voluntary side project ○ The number of people assigned to relevant tasks ○ How suitable these people are to perform the tasks
Communication	The nature of internal conversations related to the relevant activities: ○ How frequently relevant topics or tasks are discussed ○ The depth of the conversations ○ The formality of conversation channels for relevant topics. For example, is there a dedicated time to talk about it, or does it come up occasionally in Slack?

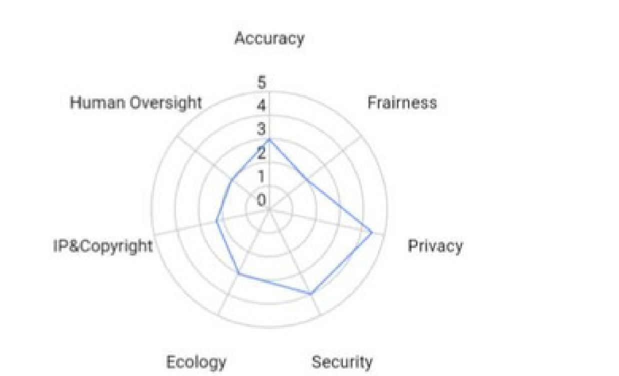
Aggregation

The maturity model allows two aggregation modes. One is based on the NIST pillars: MAP, MEASURE, MANAGE, and GOVERN. Each of these pillars represents a kind of activity, mapping impacts, measuring them, managing them, and governance activities. The statements in the questionnaire represent concepts belonging to one of them, which allows aggregating by pillar. For example, the MAP score is the average of all the statements that are based on recommendations included in the MAP pillar. This aggregation mode helps companies diagnose what kinds of activities they're strong at, and which they need to improve. Another option is to aggregate based on some or all of the dimensions of AI responsibility the RMF identifies, such as fairness and security. This aggregation mode helps companies diagnose their strengths and weaknesses with attention to specific risk areas.

Aggregation by NIST Pillar



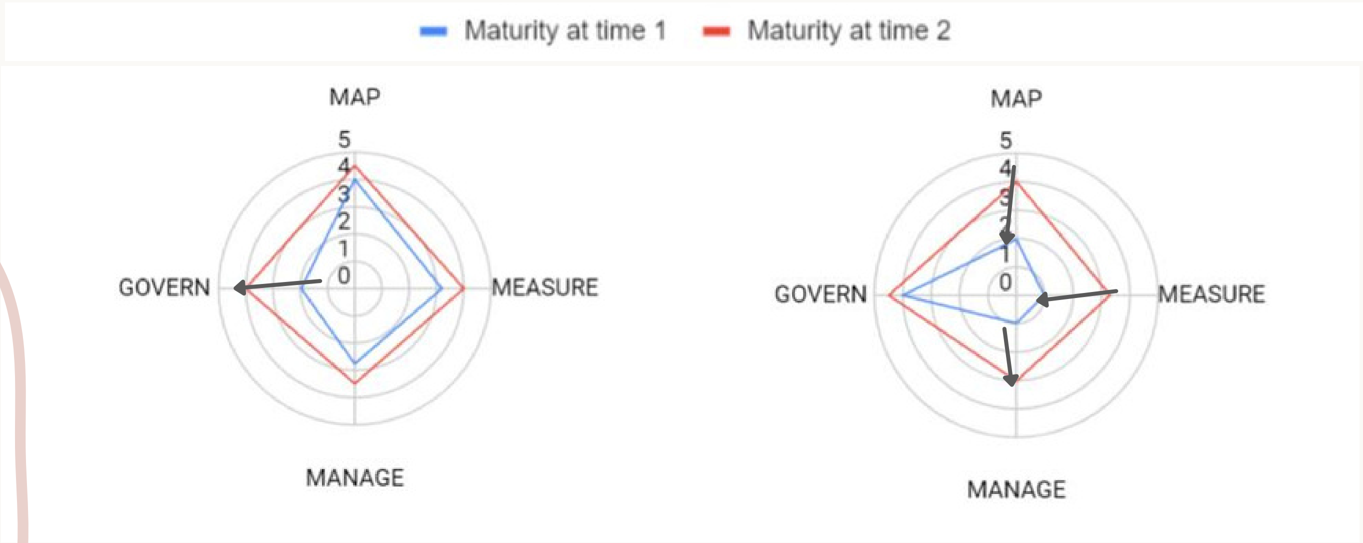
Aggregation by Responsibility Dimension



Evaluations over time reveal the company's progress. For example, top-down trajectories might begin with strong GOVERN performance and later develop to improving performance in the other pillars. Bottom-up trajectories may begin with strong MAP, MEASURE, and MANAGE performance, and later improve to better GOVERN performance.

Bottom-Up Trajectory

Top-Down Trajectory



Who does the maturity model help?

When asked who the model can help, survey respondents in the hackathon named the following: all companies (61%), auditors (24%), companies at the beginning of their AI ethics journey (22%), academics (13%), executives (11%), and consumers (11%).

Companies can use the model to evaluate themselves and strategize. External stakeholders, such as auditors, can use the model to evaluate companies as independent observers on behalf of the company or other external stakeholders, such as investors, buyers, or consumer groups. Academics can use the model for research and teaching.

All companies

61% of respondents said that companies stand to benefit from the maturity model. The usages include upskilling, evaluating themselves, and devising AI ethics strategies.

"The guidelines can help companies that are implementing/utilizing/plan to utilize AI products and services to create their own AI Ethics blueprint"

– Cigdem Patlak
Software Engineer

Auditors

24% of respondents said that the maturity model is helpful for auditors.

"Auditors and regulatory bodies can utilize the questionnaire as a standardized tool for evaluating companies' AI governance practices. It provides a structured framework for assessing compliance with relevant regulations, standards, and ethical guidelines"

– Anonymous
AI Ethicist

Early-stage Companies

22% of respondents said that the model would especially benefit companies at the beginning of their AI ethics journey.

"Small to medium-sized enterprises or startups that may not have the resources to appoint a dedicated person or team for AI governance can significantly benefit from these tools. The questionnaire and guidelines provide a structured approach to evaluating their current AI governance practices, helping these organizations identify areas of strength and weakness"

– Anastasiia Gaidashenko
Research and Policy Professional

How do evaluators benefit from the process?

The activity of filling out the questionnaire helps the evaluators themselves. When asked what they learned from the experience, the most common learnings identified by hackathon survey respondents were as follows.

How to evaluate governance maturity - 37% of respondents said they benefited from learning how to evaluate the responsibility of AI governance.

> The experience was eye-opening for me. It was very helpful to have a framework to work off of. I often think about the topics that the framework walks you through, but until working with it I did not have a robust way of assessing the topics and a way to ground and level my assessment. I found it very helpful.
>
> – Anonymous
> Software Engineer

Lack of transparency in the industry

24% of respondents said the hackathon helped them notice the lack of transparency about AI governance practices in the industry.

> “First and foremost, I learned that some AI companies (actually, probably a lot of them) do not publicly provide enough information to enable a thorough assessment of AI program maturity. Much of the publicly available content that I reviewed included lofty claims/puffery and, what seemed like, ethics-washing.”
>
> – Farrell Wilkerson
> Associate General Counsel

AI governance evaluations are important - 22% of respondents said that the experience taught them that evaluations of the social responsibility of AI governance are important.

> “I learned the importance and value of evaluating a company against a relatively empirical set of standards such that we can step away from subjective biases and narratives and lead the way towards a more constructive/nuanced conversation around a company's AI ethics.”
>
> – Devyn Greenberg
> Technology Investor

Increased AI ethics expertise - 22% of respondents said that working with the maturity model increased their expertise in AI ethics.

> “I learned a great deal from this experience. I do not have much hands-on experience evaluating companies so this is a plus...I feel like I am expanding my understanding especially since I live in the US which has a very aggressive private sector.”
>
> – Dominique Greene-Sanders
> Policy Advisor

Profiles

Select Organizers, Winners, & Participants

Ravit Dotan

Hackathon Lead
TechBetter, CEO

“The hackathon was a wonderful experience for me! Most of all, I enjoyed interacting with the participants in all the different activities—the online sessions, the office hours, the Slack channel, and the assignments. The levels of engagement and enthusiasm were high, and so was the quality of the comments and work the participants produced.

We, the team behind the hackathon and maturity model, have learned so much from the participants. One of the main reasons we organized this hackathon was to bring diverse perspectives into the development of the model. So many frameworks are created by small teams or ingest input in narrow ways. We wanted to do things differently. We wanted to embed diverse perspectives deeply in the model’s design. The participants’ thoughtful comments allow us to do just that. For example, one of the most important learnings is how to make the model more accessible for smaller companies. Many standards, including the NIST AI RMF, have been criticized for favoring large organizations. We have tried to address this problem in the questionnaire from the get-go, and the engagement with the participants taught us new and better ways to increase the questionnaire’s accessibility.

Moreover, the power of hackathons as a tool for engaged research and education became clearer to me than ever. Traditionally, research and education are separate activities, and, often, education is a largely passive process with limited impacts that go beyond the learning itself. For example, in universities, undergraduate students typically learn by attending lectures and writing assignments with limited external impact. Research is a separate activity conducted by others. The hackathon was an experiment with a different model: Learning by doing that facilitates research. Many hackathon participants have indicated in their feedback forms that the hackathon helped them upskill in AI ethics. That upskilling happened through engaged activities – conversations, communal brainstorming, and assignments. The outputs of the activities contributed to something beyond the learning process itself: research on the maturity model. This interaction between education and research was by design. Seeing it unfold was spectacular.

I look forward to many more events of this kind!”

Naagma Timakondu

Participant

“I am a cognitive scientist who is passionate about using an understanding of the human mind and behavior to design value-aligned technology. I have a master's in Applied Cognition & Neuroscience and bachelor's in Cognitive Science. Prior to obtaining my master's, I had the privilege of working in IBM's AI Ethics Project Office supporting IBM's AI Ethics Board. Through this experience, I worked on projects to help implement AI principles into practice, building tooling to support employee engagement, and cultivating a community to champion a culture of AI ethics…I learned how the NIST AI Risk Management Framework works and how to apply it when evaluating an organization's AI practices. I learned how easy it is for organizations to claim they have principles in place but difficult for them to provide public documentation on how they are putting those principles into practice. I learned how to comb through public documentation to extract concrete evidence on the AI maturity of an organization [rather] than base it off of general principles.”

Borhane Blili-Hamelin

Organizer

“The hackathon challenged my thinking about the place of our maturity model in the AI risk management space. To me, we must work towards virtuous feedback between standards defining "good enough" governance and frameworks aimed at growth.

Compliance frameworks often ask a binary question. Does an organization implement good enough practices to meet the expectations of regulators or standards? Work with colleagues at BABL AI shows the power of criteria-based bias audits for such questions.

I love this quote from Gregory Fontenot, cited in Micah Zenko’s book on red-teaming: “When you hear “best practices”, run for your lives. The Titanic was built with best practices. It was faithfully operated in accordance with best practices.” I see practices like red-teaming as sitting on the other end of the spectrum: when organizations worry that what looks good enough might be deeply harmful, red-teaming can help.

Similarly, the NIST AI RMF focuses on growth by prompting organizations to ask comprehensive questions about improving their AI risk management.

Before the hackathon, I saw our maturity model as similarly growth-oriented. The hackathon helped me appreciate its potential in bridging the gap between good-enough and growth-oriented frameworks. To make improvements to our model that tangibly meet the needs identified by participants — such as more tangible scoring baselines, examples, and guidelines, or greater consistency — we need to more explicitly tackle where “good-enough” governance falls on our maturity model. More generally, the hackathon helped me appreciate how bridging this gap is an unmet need our model is well-positioned to tackle.

On a more personal note, I was astounded by the All Tech Is Human community's level of engagement. I’ve experienced and designed a fair share of participatory online workshops. But until the January 31 workshop, I had never seen a group of 170+ participants so passionately engage in 90 minutes virtual collaboration.”

Philippe Schroeder

Honorable Mention

“This experience deepened my understanding of Anthropic and the challenges in responsible AI practices within corporate landscapes. I learned that the scarcity of trustworthy third-party and publicly available information complicates evaluations, requiring heavy reliance on self-reported company data. I also found that companies often present philosophical ideals, but their translation into tangible AI implementations in their models remains opaque. Extracting detailed insights for applying the NIST framework, even from well-documented companies, is hard. Additionally, I think it is hard to evaluate a company without one's judgment being influenced by the knowledge of other companies and then comparing it to them...I believe the questionnaire helps companies to ensure they design their AI systems as responsibly as possible. It provides accountability, can unveil weaknesses, and serves as a guiding document for companies who want to do better but don't know how. “

Jeanna Matthews

Organizer

“I was delighted to see the level of interest in participating in the hackathon. There were over [180] people at the first Maturity Model workshop in January which was exciting. Some people came with an intention of helping their company or organization do a first maturity assessment. Some people came with an intention of developing their ability to offer maturity assessments. Some people came to give feedback on and help us improve our maturity model. Over 50 people submitted entries to the hackathon which is a sizeable portion of those attending the first workshop. People who participated reported increasing their AI ethics expertise and gaining appreciation for the lack of transparency too often found in industry, as well as learning how to evaluate maturity with respect to AI risk management. All of those are great outcomes! As a computer science professor, it was especially gratifying to me to see interest among current students and even a former student.”

Katherine Grillaert

Winner

“[A]nalysing the signals was surprisingly resource intensive. I found that the company I assessed used a lot of jargon and keywords, and made broadly agreeable statements. However, they were lacking in substance and specifics. I wanted to be fair, so I looked very hard for other supporting documentation for my company. On the other hand, part of responsibility and transparency is that this information should be straightforward to find. I could see this being helpful for companies who want a clear pathway for creating their responsible AI approach and documentation, as it translates the NIST AI RMF to actions.”

”I could also see this being used to assess companies and provide consumers with a snapshot based on aggregate scores from several reviewers. Based on my experience evaluating, it would be quite difficult for consumers to understand a company's responsible AI maturity, even if the public facing documentation is written in plain language. This assessment can provide a snapshot that is a useful tool for consumers.”

Cigdem Patlak

Participant

“The adoption of AI technologies is taking place at such a rapid pace that for an outsider it is a quite unrealistic goal to keep up with all the news and track down all the risks that may arise upon deployment and broader use of AI in a wide variety of industries. Companies need to proactively invest time and money into self-regulation efforts rooted in Responsible AI practices, ideally such practices should be based on multi-disciplinary collaboration and research of industry, academia, civil society and governments in order to have common ground on trustworthy AI policies and eventually laws.”

“The questionnaire provides an opportunity to conduct a thorough evaluation of a company’s technological infrastructure and its alignment with Responsible AI principles. It offers practical guidance to set up ongoing evaluation efforts to ensure responsible technology principles are upheld in line with the progression of AI technologies. Many ethical, legal and social risk may not have been identified and there is a gap in laws and policies to address these issues as of now. As a result, this questionnaire is a companion to revisit and reflect on at regular intervals to define new actionable steps and it may even need to evolve with further/modified questions, if the need arises.“

Andrew McAdams

Participant

“The questionnaire and guidelines help internally experienced professionals and those developing AI Governance programs the best. For experienced professionals, it helps translate into what processes and policies they are already familiar with to satisfy the various questions about a Responsible AI program. For those developing AI Governance and Responsible AI programs, I think it provides a roadmap on the types of controls they need to evaluate and implement to help achieve their targeted degree of maturity. However, I don't think questionnaires and guidelines can be picked up by someone without an understanding of software development, policy, support, etc. There was too much through the few things that I answered where I knew what they were getting at without them explicitly [stating] it, and so there was a lot of me reading between the lines, making assumptions, and looking for implications. ...I don't think the questionnaire works as well for the current state of disclosure of AI Governance systems. Some of the questions and sub-statements ask for fairly detailed explanations of internal processes and procedures that could open the company up to increased risk because of the public disclosure. I was able to muddle through in most cases, but I also scored every metric lower because I couldn't verify or even "read the tea leaves" well enough to determine if Inworld satisfied the requirements of the sub-statements.”

Anastasiia Gaidashenko

Winner

“Two key lessons stand out: 1. The crucial role of thorough and systematic documentation in AI governance. Meta, like many large tech companies, undertakes numerous initiatives to manage AI risks. However, the evaluation revealed that even a company as advanced as Meta could improve by making their documentation more comprehensive and systematic. This includes explicitly detailing plans for transferring or avoiding risks, and providing more information on how responses to risks are prioritized based on impact, likelihood, available resources, and the organization's risk tolerance. 2. Inherent difficulty of evaluating a company's AI governance practices from the outside. The abundance of information, much of which may be irrelevant or not directly related to AI governance, poses a significant challenge. It highlights the need for clear, accessible, and targeted communication from companies about their AI governance and risk management efforts. For external evaluators, it underscores the importance of developing sophisticated techniques for sifting through vast amounts of information to extract relevant insights. This difficulty also points to the potential value of standardized reporting or disclosure frameworks for AI governance, which could facilitate more effective and efficient external evaluations.”

“Journalists can use the questionnaire and guidelines as a framework to investigate incidents involving AI systems. By understanding the components of robust AI governance—such as documentation of risks, response plans, and measures of input diversity—journalists can better analyze the sources of failures or ethical lapses in AI applications. This structured approach can help uncover not just the symptoms but the underlying governance weaknesses that lead to problematic incidents.”

Dr. Joshua Scarpino

Organizer

“The AI maturity hackathon underscored the importance of a maturity model that aligns with industry-standard frameworks. It also highlighted the need for foundational training and education for individuals involved in the responsible AI ecosystem and raised awareness of the current skill gaps. Assessing maturity within an organization is important for a couple of reasons. Internally, this helps the organization plan and establish strategic roadmaps while ensuring alignment with internal risk tolerances. Externally, this can help drive and support conversations for investment decisions, build consumer trust, and validate compliance with relevant regulations, standards, and ethical guidelines.

Organizations often face challenges in adopting appropriate AI governance and practices. The AI risk maturity model is crucial for organizations seeking to understand, assess, and improve their programs and processes. It provides a North Star, guiding organizations in their AI maturity journey and paving the way for continuous improvement.

Dr. Joshua Scarpino

The establishment of a standard benchmark for organizations' maturity is pivotal in promoting transparency and confidence in the technologies they deploy. This method allows organizations to evaluate their progress against their goals and peers, fostering a culture of continuous learning and improvement. Increasing education and awareness across the AI ecosystem is another critical focus area. There is a fundamental need to ensure that all individuals who are part of the design, deployment, use, and assessment of AI systems and processes are appropriately educated and trained. This ensures appropriate identification of risks and raises awareness around ethical challenges, ensuring that harm is mitigated and not unintentionally perpetuated at scale. Ensuring technology systems and supporting processes are well explained and that individuals understand a given system's purpose, design, and impact is critical; this is possible and can be accomplished through continued evaluation and transparency.

Overall, the Hackathon was highly successful in solidifying the need for a Responsible AI Governance Maturity Model and highlighting the need for continued education. Continued refinement of this model will benefit companies across all sectors and provide a standard approach to understanding the level of maturity within organizations.”

Amari Cowan

Participant

"I am an AI Governance Fellow at the Portulans Institute, where I study existing and theoretical frameworks for the governance of general purpose AI in society. Before this role, I worked on machine learning policies at Meta, where I helped to craft explainability materials, build novel governance frameworks, and address risks related to emerging technology at Meta...[This] can certainly help industry professionals who want to improve AI systems in good faith, but also improve user's understanding of how they should perceive or measure a 'good' or safe AI policy. Users deserve to have access to documentation that allows them to make informed decisions about the AI-enabled products they choose to use."

Alicia Rémont Ospina

Honorable Mention

“I am the Risk Partner for the teams working on developing products embedding features leveraging Artificial Intelligence at Swift. I am also contributing to the development of Swift's AI Governance framework as the main point of contact for Enterprise Risk Management and to create links between existing frameworks and methodologies and new Responsible AI principles that the company is looking to adopt.”

“[T]his is a great questionnaire for Ethical AI leaders trying to build and enforce frameworks in their companies as it provides a rather comprehensive list of elements to keep in mind when applying Responsible AI standards to the development of AI systems. It could also be condensed into a sort of to-do list for engineers and product owners/managers for them to keep in mind ethical principles when developing AI systems.”

Monika Viktrova and John Walker

Participants

“I [Monika Viktrova] partnered with John Walker to jointly evaluate Inflection's Pi.ai using two different approaches both based on the same principle: can LLMs evaluate LLMs effectively, given a set of clear instructions?”

“Awareness of AI ethics in the public has grown. Nowhere is this more clear than in startups having pages on “Safety”. These pages include information not just on the basics of privacy, which would be mandated by GDPR before a tool can be opened to users in the European Union, but on extra points around risk management and ‘values based development’. However, much like their LLM’s responses, these pages are often vague, high-level, and provide little ability for a given user to verify their claims. The lack of transparency calls into question the ability to trust the bulleted lists and raises the question of whether they are meaningfully more than ethics washing.”

“It's clear that the maturity model will be more effectively leveraged in partnership with the tech teams building or deploying a given system. Because companies rarely reveal the inner workings of their systems and indeed consider them trade secrets, and because no legislation is currently in force to ensure this transparency, public documentation about AI systems is scant and high-level. Only with the help of those who understand, maintain and scale a given system can maturity truly be evaluated.”

Lindsey Washburn

Participant

“I had a few companies in mind to try for the evaluation, but had a similar issue with all of them - they didn't have much public data available. I thought about doing the evaluation for a larger, publicly traded-company, but I liked the idea of doing a deep dive on a company that my company is working with, especially since responsible AI is something I personally care and am passionate about. The evaluation showed me what components are important for a company to have in place and will help me ask the right questions going forward as my company continues to work on this project with them.”

Amy Daley

Honorable Mention

“I believe the questionnaire and guidelines are helpful mostly for internal use by companies or as partners with an ethics group. The maturity model gives them guidelines of areas to include on their websites to provide clarity to the public that they are working to implement AI ethically and responsibly. Because the information is not always posted to the public, I think it's difficult to provide accurate feedback to companies without their involvement...I rather accidentally picked a company with lots of information on their website.”

Caroline Lancelot Miltgen

Honorable Mention

“The questionnaire is quite long and sometimes hard to answer. It may be because the organization I assessed does not say that much about responsible AI. What was particularly difficult was that the organization did not put everything in a single and complete document. Hence, you have to check all over the place to find information about their process to be able to answer the questions.”

Fouzia Ahmad

Participant

“If a company has more than one AI product, should there be a maturity model evaluation for each product, rather than a company-level evaluation? Presumably, a very mature organization will have the same processes and characteristics for all of their products across the board, but the reverse is not true. A non-mature org can have different levels of maturity for their different products...Some major initiatives done by an AI company, which may not qualify as a typical “product” (for example, Scale AI’s SEAL initiative, which is an “evaluation product”), may need some different sub-statements in order to better evaluate them.”

Carol Anderson

Organizer

It was exciting to see so many participants, from such a broad range of backgrounds, join the hackathon. I think this speaks to a great hunger for tools that can help people operationalize the principles of Responsible AI. It was also great to see participants apply the Responsible AI Governance Maturity Model to such a broad range of companies in their hackathon entries. Their work brought to light a number of great ideas for improving the framework.

Case Study: **Light-It**

Case Study: Light-It

Light-It is a startup that used the maturity model to evaluate itself, and they discussed their experiences and learnings with hackathon participants through a panel.

Here we summarize highlights from the panel.

About Light-It

Light-It is a digital product agency building tailor-made healthcare web and mobile applications. They partner with digital health companies, healthcare innovation centers, and startups to reach technology's full potential by ideating, designing, and developing custom applications that revolutionize the industry.

Puppeteer, a subsidiary of Light-It, is a platform that enables healthcare organizations to leverage generative AI seamlessly, safely, and in a compliant way. It aims to revolutionize healthcare with Conversational AI, creating intelligent, human-like LLM-based healthcare applications for maximum efficiency.

Who participated?

Adam Mallát, an innovation manager, and Javier Lampers, founder and CEO, evaluated Light-It using the maturity model and discussed reflections at the hackathon.

Javier Lampert
Founder & CTO

Adam Mallát
Innovation Manager

Why do you care about AI responsibility?

The Competitive Advantage of AI Responsibility

Light-It stays ahead of competitors, including much larger ones, thanks to its expertise in addressing safety and compliance issues. In developing AI-based applications, the competitive advantage is even larger relative to other software. Since there are no strict AI-specific regulations and guidelines yet, many competitors are still far behind on AI responsibility, so that creates a considerable opportunity for a competitive edge. Moreover, AI responsibility is key to ensuring compliance with sector-specific regulations, such as HIPAA.

> "We managed to to have a competitive edge by being experts at safety and compliance."
>
> – Adam Mallát

The Ethical Value of AI Responsibility

Having a positive social impact is important to the company. In the healthcare sector, the stakes are especially high. For example, when patients are communicating with chatbots for mental healthcare needs, lives may be at stake. Therefore, for example, when developing such a chatbot, Puppeteer ensures that if the chatbot identifies suicidal or self-harm thoughts, the chat stops and the person is immediately referred to a human care provider.

What did the evaluation process look like?

The first step was meeting with the hackathon team to learn about the maturity model. Then, Adam, the innovation manager, filled out the questionnaire, and Javier, the CTO, reviewed it. The last step was another meeting with members of the hackathon team, in which we went over the evaluation and discussed learnings from it.

How did you explain scores?

Large companies often create a lot of documentation and structured processes, which may then be used to show the company's priorities. However, as a startup, Light-It has less of an emphasis on documentation. Light-It uses an agile approach, which allows them to be nimble and have a lot of control over the product. Their prioritization is reflected in the company's objectives and resource allocation. For example, how many employees are empowered to work on the topic? Is the topic reflected in their Objectives and Key Results (OKRs)? Do they track measurable metrics related to the topic?

Did the questionnaire help you?

Engaging with the questionnaire became an opportunity to think systematically and analytically about the company’s efforts in AI ethics.

> “[Filling out the questionnaire] was the first time, that I personally went into so much detail and became so analytical about what we do in the area of of ethics”
>
> – Adam Mallát

Did you identify AI responsibility growth opportunities?

Risk Assessment and Mitigation

The questionnaire helped the company think carefully about some of the risks related to their products, such as bias. In particular, it helped them understand that all AI systems face bias risks and they have decided to to further empower their engineers to identify and reduce these risks.

> “Bias is one of the main points we are going to tackle in the future that we weren’t...We are now developing tools to allow developers to understand bias”
>
> – Javier Lampert

Documentation

In addition, Light-It is considering adding documentation related to AI responsibility. Their main priorities are documents related to risk assessment and risk management.

About
The Hackathon

About the Hackathon: What did we do and why?

Hackathon goals

The hackathon had two goals:

- **Participatory development** - The hackathon was a way to include a large group in the design of a product to help the community, as we used learnings from the hackathon to improve the model.
- **Mutually beneficial feedback collection and engaged learning** - we aimed to make the hackathon a mutually beneficial experience and to facilitate active learning. Therefore, the hackathon included educational components, supporting participants in their educational journey in AI ethics and AI governance. Moreover, the educational components involved active assignments rather than passive content consumption.

Hackathon activities and tools

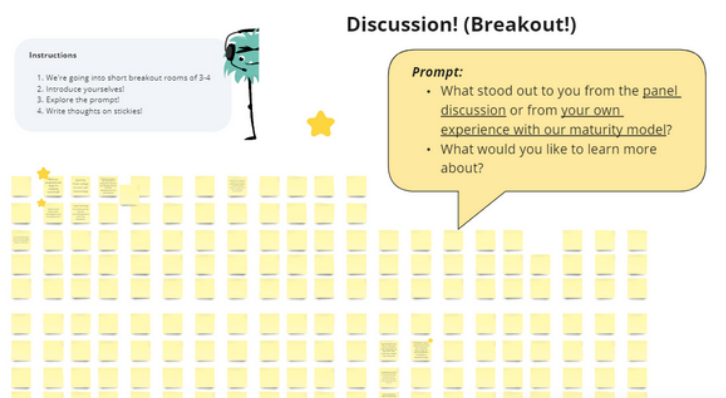
The hackathon was fully virtual. It included **two interactive online sessions**, **three office hours**, and an **asynchronous assignment**.

We used the following tools and activities:

- Mini-talks, a panel, and group discussions
- Breakout rooms
- Online form for the assignment
- Slack channel for ongoing communication
- A Miro board for group brainstorming (see screenshot below)

To the right:

A part of the board we used for group brainstorming.



Discussion! (Breakout!)

Instructions

1. We're going into short breakout rooms of 3-4
2. Introduce yourselves!
3. Explore the prompt!
4. Write thoughts on stickies!

Prompt:

- What stood out to you from the panel discussion or from your own experience with our maturity model?
- What would you like to learn more about?

Online sessions

First online session (Jan 31, 2024)

In the kick-off session, participants learned, through participatory lectures and Q&A sessions, about the basics of AI governance and the AI maturity model. We also facilitated a number of activities examining the risks of ChatGPT — including a structured large-group activity, and a small-group activity in breakout rooms — to promote active learning in connecting AI governance questions to a concrete use case. We ended by describing the hackathon assignment.

Second online session (Feb 7, 2024)

We began with small group check-ins between participants to discuss their experiences using the maturity model so far. We then held a panel discussion with employees of Light-It, an AI startup, about their process of applying the maturity model to their own business. This was followed by a Q&A session where participants asked questions of the panelists and hackathon organizers. We ended with a structured large-group exercise in which we gathered feedback from participants about the maturity model.

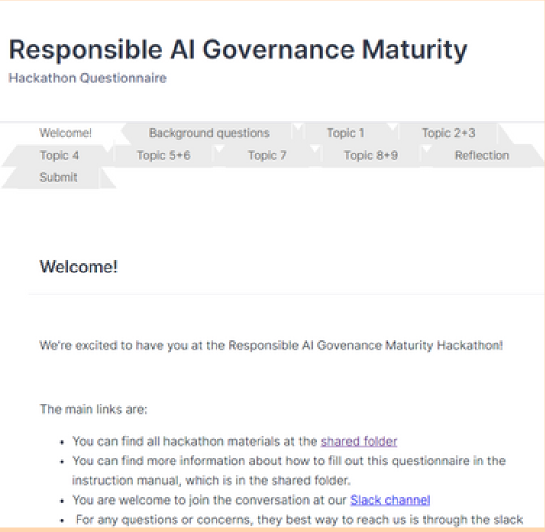
Office hours

We held three office hours, two in between the sessions and one after. Participants who came asked questions and we had free-flowing conversation.

Assignment (due Feb 10, 2024)

The hackathon assignment was to fill out the maturity model questionnaire. Participants worked individually or in small groups. Each person or group chose a company to evaluate. They either evaluated their own company based on internal information or any company of their choosing based on public information.

To the right:
A part of the online form participants used to conduct the evaluation.



Responsible AI Governance Maturity

Hackathon Questionnaire

Welcome!

We're excited to have you at the Responsible AI Govenance Maturity Hackathon!

The main links are:

- You can find all hackathon materials at the shared folder
- You can find more information about how to fill out this questionnaire in the instruction manual, which is in the shared folder.
- You are welcome to join the conversation at our Slack channel
- For any questions or concerns, they best way to reach us is through the slack

Assignment content

In the assignment, participants evaluated their chosen company on at least one topic. For each chosen topic, participants answered the same questions about each of the metrics:

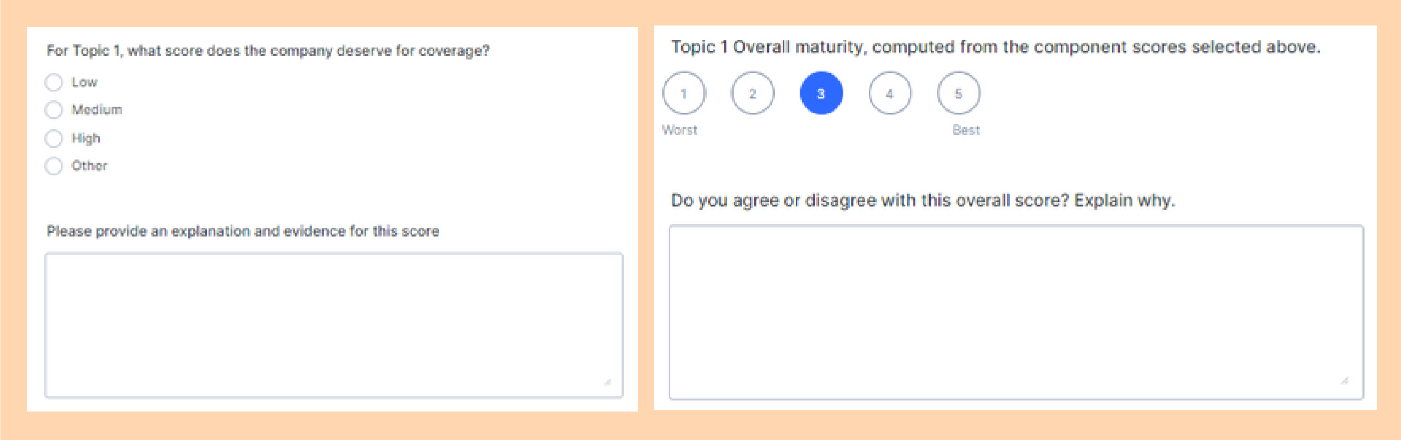
- For topic x, what score does the company deserve for **coverage**?
  - Choice between High / Medium / Low / Other
- Please provide an explanation and evidence for this score
  - Text box

- For topic x, what score does the company deserve for **robustness**?
  - Choice between High / Medium / Low / Other
- Please provide an explanation and evidence for this score
  - Text box

- For topic x, what score does the company deserve for **input diversity**?
  - Choice between High / Medium / Low / Other
- Please provide an explanation and evidence for this score
  - Text box

  The form calculated the **overall score** automatically based on the scores of all the metrics. Participants were asked:
- Do you agree or disagree with the overall score? Explain why.

Below are screenshots from the form for illustration.



For Topic 1, what score does the company deserve for coverage?

○ Low
○ Medium
○ High
○ Other

Please provide an explanation and evidence for this score

Topic 1 Overall maturity, computed from the component scores selected above.

○ 1 ○ 2 ● 3 ○ 4 ○ 5
Worst ... Best

Do you agree or disagree with this overall score? Explain why.

Reflection questions

In addition to evaluating companies, participants answered reflection questions:

- What did you learn from the experience?
- Who can the questionnaire help and how?
- How can we improve the questionnaire and scoring guidelines?
- Help us improve - Do you have any comments or feedback about the hackathon?

About the Hackathon: **Who Participated?**

Participant Backgrounds

180 Participated in at least one hackathon activity
54 participants submitted a filled-out questionnaire.
46 agreed to use their data for this research, their backgrounds are summarized below.

Roles:
- Most of the participants were in **technical roles (28%)**.
- Many were in advisory roles (22%), and governance and compliance roles (20%). Other roles include AI ethics (15%), research (15%), and product management (7%).

Industries:
- The largest group of participants was employed in **big tech (22%)**
- Some (15%) were in tech companies that are not big tech.
- The most represented sectors were healthcare (9%) and finance (9%)

Genders:
- **61% identified as women.**
- 20% Identified as men.
- The rest didn't specify

Geographies:
- The largest group identified as residing in **North America (39%)**
- Others resided in Pacific Asia (13%), Europe (7%), and South Asia (4%)

Ages:
Of the 28 participants who reported their age, the average age was **40.2**

Evaluated Companies

The following is a list of the companies participants evaluated, excluding companies evaluated by participants who didn't want to share their data. Some were evaluated by multiple participants.

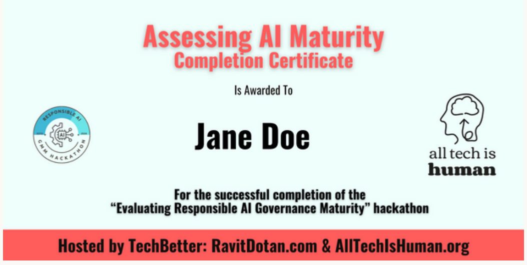
Adobe	Cohere	Google	HUL	Khan Academy	Microsoft	Patreon	Scale AI	The New York Times	Woebot Health
Anthropic	Deutsche Telekom	Grammarly	Inflection	Match Group	Nurdle AI	Rolls Royce	Snap Inc	TikTok	Workday
Checkr	Duolingo	Hirevue	Inworld	Meta	OpenAI	Savia	Stripe Radar	UiPath	YouTube

About the Hackathon:
Awards and Rewards

Participation rewards

Participants who submitted completed evaluations received:

- Completion certificates (example to the right)
- Branded swag
- Seven of the participants who completed evaluations received a 1-1 mentoring session with Rebekah Tweed or Ravit Dotan.



Assessing AI Maturity
Completion Certificate

Is Awarded To

Jane Doe

all tech is human

For the successful completion of the "Evaluating Responsible AI Governance Maturity" hackathon

Hosted by TechBetter: RavitDotan.com & AllTechIsHuman.org

Winners

Winners were chosen based on the quality of the submission. In addition to participation rewards, winners received:

- A trophy
- Mentions in All Tech Is Human announcements
- A completion certificate with an indication they won

Honorable mentions

In addition to winners, the hackathon committee also chose two honorable mentions. They were chosen based on the quality of their work. In addition to participation rewards, honorable mentions received:

- A completion certificate with an indication they received an honorable mention
- Mentions in All Tech Is Human announcements

About the Hackathon: **Winners and Honorable Mentions**

Winners

In alphabetical order, the three winners are:

Narcisa Codreanu
Anastasiia Gaidashenko
Katherine Grillaert

Honorable Mentions

In alphabetical order, the honorable mentions are:

Amy Daley
Caroline Lancelot
Alicia Rémont Ospina
Philippe Schroeder

About the Hackathon: **Participant Experiences**

"The experience was eye-opening for me. It was very helpful to have a framework to work off of. I often think about the topics that the framework walks you through, but until working with it I did not have a robust way of assessing the topics and a way to ground and level my assessment. I found it very helpful".

– Anonymous,
Software Developer

"The framework is a great sounding board for organisations at different phases of their AI journey. It enables them to catch issue[s] early in the system reducing cost and reputation implications"

– Payal Padhy,
Technology Consultant

"I think that the AI Governance Maturity Model raises many excellent questions that AI companies and projects should strive to address, and to do so in a requisitely comprehensive way. I value having the Maturity Model as a reference to return to as our project continues to develop, and as a tool to plan for future considerations."

– Jeffrey Perrone,
UX Designer & Tech Strategist

"It was great to see a public hackathon in this space, and also get the chance to meet such a diverse group of participants. The multiple office hour sessions at different times of the day were also a useful addition for folks across timezones and with different availability schedules."

– Yohan N.V. Mathew,
Data Scientist & Software Developer

About the Hackathon: **Participant Experiences**

"The hackathon is an amazing opportunity for people from all walks of like to learn more about AI and its allied technologies."

– Arijit Goswami,
Senior Innovation Manager

"It was a great learning experience. As far as I'm concerned, that experience is what I was yearning for."

– Christopher Chitimbwa,
Software Developer

"I really liked how friendly and accommodating the team was about the hackathon, as well as the efforts to ensure the training was engaging."

– Alyssa Png,
Lawyer

"I like the system and think it is helpful"

– Maria Tran,
Lawyer

"I loved the breakout room and using Miro for an interactive experience! ...

I appreciate the effort from the whole team and thank you for the opportunity!!"

– Naagma Timakondu,
Cognitive Scientist

"This was a very interesting project. It seemed simple and straightforward at first, which is a testament to the team's hard work at preparing the maturity model assessment from the NIST AI RMF"

– Katherine Grillaert,
Interdisciplinarian

"I loved the diversity of thoughts that were gathered on the Miro board."

– Philippe Schroeder,
Product Designer

"This was a fun and useful exercise and the session and office hours around the project was also great! Thank you for hosting this project."

– Shasti Walsh,
Product Manager

"[I learned] the high value of a well-defined matrix that can assess, measure, and guide the development of AI"

– Deborah Hagar,
Sustainability and AI ethics

From the Hackathon: **Key Learnings**

From the Hackathon:
Key Learnings

1. What do evaluators aim to do?

Participants aimed to determine whether companies engage in good or bad behaviors, or at least identify what they don't know:

- **Evaluate good behavior**
  - Through evidence that the company **implements (some) responsible** AI practices.
- **Evaluate bad behavior**
  - Through evidence that the company **has not adopted (some) responsible** AI practices (e.g., no company resources are used for AI responsibility).
  - Through evidence that the company is **implementing (some) irresponsible** AI practices - for example, lawsuits or internal documents indicating that the company uses discriminatory methods.
- **Identify unknowns**
  - Articulating when one can't tell what the company does due to a lack of evidence.

2. What information do evaluators use?

Participants looked for the following types of information to evaluate companies' behaviors:

- **Execution –** Outcomes, procedures, and resources dedicated to activities:
  - RAI metrics and progress on those metrics
  - How much time is spent working on certain tasks
  - The execution of RAI best practices, such as red-team exercises or ethics reviews

- **Uptake** - How the relevant activities and deliverables are received:
  - Whether the outputs of RAI work are officially adopted by the company
  - Whether and how the company's leadership supports it

- **People** - Who conducts the relevant activities and how:
  - Whether the people conducting the relevant activities are doing so as part of their official capacity or as a voluntary side project
  - The number of people assigned to relevant tasks
  - How suitable these people are to perform the tasks

- **Communication** - The nature of internal conversations related to the relevant activities:
  - How frequently relevant topics or tasks are discussed
  - The depth of the conversations
  - The formality of conversation channels for relevant topics. For example, is there a dedicated time to talk about it, or does it come up occasionally in Slack?

From the Hackathon: **Key Learnings**

3. Where do evaluators look for information?

Participants used different sources to find relevant information depending on whether they evaluated the company externally or internally:

External sources of evidence

- External company documents, such as their AI ethics frameworks
- External company reports, such as annual or ESG reports
- Research papers produced by the company or about the company
- Media reports
- Lawsuits

Internal sources of evidence

- Internal documents
- Interviews with employees
- Informal conversations and employee knowledge
- Internal metrics - e.g., Objectives and Key Results (OKRs) and Key Performance Indicators (KPIs)

4. Documentation requirements disfavor small companies

The questionnaire we used in the hackathon highlighted documentation as a key kind of evidence. For example, it asked whether the evaluated company conducts certain activities *and* documents the process. This choice followed the NIST AI RMF, which emphasizes documentation.

Hackathon participants pointed out that emphasizing documentation may disfavor smaller companies, whose work style is often informal. Instead, participants suggested using indicators related to the company’s priorities, resource allocation, execution of best practices, etc.

From the Hackathon: **Key Learnings**

5. Scoring Dilemmas

Dilemma 1: How should a lack of evidence impact the score?

Participants sometimes didn't find any evidence regarding the topic they were evaluating, especially when they were evaluating the company externally. When participants didn't find any information regarding a certain metric, they usually marked the score for that metric as "Low," which decreased the overall score of the company. Some participants were happy with this point deduction for lack of evidence, but others disagreed, arguing that a lack of evidence shouldn't decrease the company's score.

Dilemma 2: Should there be negative scores?

A choice between Low, Medium, and High doesn't allow participants to express more negative judgments of companies, which may arise when there is strong evidence of negative behavior. In the hackathon, participants didn't use evidence of bad behavior much. When they did, it was related to lawsuits or negative media coverage. In these cases, the score they chose reflected a balance between the good and bad they found. An open question remains: What if the bad outweighs the good? Should there be negative scores?

Dilemma 3: How to balance between different sub-statements?

When evaluating companies on a given topic, sometimes the company's performance in certain sub-statements was much stronger than their performance in others. For example, in Topic 1, mapping risks, companies may be very clear about the business value of the AI system but not about the potential risks. In such cases, participants sometimes overlooked the negative or the positive performance. One potential solution to this difficulty is to allow evaluators to rank each sub-statement separately. As many evaluators discussed each sub-statement separately in their explanation anyway, this solution may keep the workload at a similar level.

Dilemma 4: Relative or absolute scoring?

Evaluations differed on whether they ranked companies on relative or absolute scales. Relative evaluations compare the company's performance to others, so companies would get high scores if they did better than others in their areas. Absolute evaluations compare the companies to an ideal or to a list of activities they could perform. In these cases, companies' scores depended on their actions regardless of what the competitors were doing.

From the Hackathon: **Key Learnings**

6. Scoring Granularity

Evaluators tended to give companies the score of “Medium.” This makes it difficult to differentiate between companies clearly. One solution to this issue is to increase the scoring granularity, adding more levels to the existing “Low”, “Medium”, and “High” ranks.

7. Evaluating Whole Companies

In the research paper that grounds the questionnaire, evaluators may choose between evaluating AI governance at a company as a whole and evaluating the governance of particular AI systems. In the hackathon questionnaire, evaluators were asked to evaluate companies as a whole. We made this decision to simplify the evaluation process. However, the evaluation revealed that evaluating whole companies, especially large companies, is not granular enough.

Participants indicated that they would have benefited from additional training about the questionnaire. They proposed two kinds of training that could empower them:

Examples

Filled-out questionnaires and sample questions would help evaluators get a sense of what kinds of things count as evidence, what to say in the explanation of each score, and so on.

Quizzes

Quizzes could facilitate comparisons to the scoring of previous quiz-takers or to sample scores. This kind of training would help evaluators align on when each score is appropriate.

Hackathon Excerpts:
Exemplary Evaluations

Topic 1: Mapping Impacts

1. **We clearly define what the AI is supposed to do and its impacts**, including scope, goals, methods, and negative and positive potential impacts of these activities:

1.1 We define the **goals, scope, and methods** of this AI system.
1.2 We identify the **benefits and potential positive impacts** of this AI system, including the likelihood and magnitude.
1.3 We identify the **business value** of this AI system.
1.4 We identify the possible **negative impacts** of this AI system, including the likelihood and magnitude.
1.5 We identify the potential **costs of malfunctions** of this AI system, including non-monetary costs such as decreased trustworthiness.
1.6 We implement processes to integrate input about **unexpected impacts**
1.7 We identify the **methods and tools** we use for mapping impacts.

Coverage

Amy Daley - Webot Health, High

I easily found discussion of the first three substatements (A, B, C) on the homepage (woebothealth.com) as well as on the Technology Overview page (https://woebothealth.com/what-powers-woebot/). In substatement A, they do a good job of outlining goals. They are offering a framework to companies and an app to individuals that allow mental health therapeutic conversations using three evidence-based treatments CBT, DBT, IPT). Part of the goal is to expand outreach to those who don't have access to human mental healthcare and also to allow companies to integrate these tools into existing clinics. Their scope is also articulated as four areas: adults, adolescents, maternal mental health (all currently deployed) and a fourth area for substance use in development. Methods are a little harder to find articulated, but available in several linked articles and a blog post. For example, they are currently offering a rules-based conversational agent using natural language processing, but they are also assessing the risks of moving into the implementation of a LLM and a true generative AI.

For downsides (D, E, F, G) there is discussion but not necessarily meeting the needs of the maturity model. For example, for Substatement D some of the possible negative concerns are discussed in the AI Core principles https://woebothealth.com/ai-core-principles/. Substatement E: costs of potential malfunctions is discussed at a high level on the same page. For Substatement F: implementations of processes for unexpected impacts they note they undergo regularly external security certifications (SOC 2 Type 2 + HIPAA Compliance Report and ORCHA DHF are two notable ones mentioned on https://woebothealth.com/safety/. web page). Substatement G: documentation of methods and tools for mapping impacts, they don't seem to have this as obviously outlined, but I suspect they have this internally.

Sources:

- https://woebothealth.com/what-powers-woebot/
- https://woebothealth.com/ai-core-principles/
- https://woebothealth.com/safety/

Topic 2: Identify Requirements

2. We identify the requirements the AI must meet, including compliance, certifications, and human oversight needs:

2.1 We document the **human oversight** processes the system needs.
2.2 We document the technical standards and **certifications** the system will need to satisfy.
2.3 We document AI **legal requirements** that apply to this AI system.

Coverage

Caroline Lancelot - Grammarly, Medium

While the organization documents human oversight ("human in the loop, ensuring user autonomy, we put users in control of their experience"), they also document technical standards and certifications (they say "See the attestations and certifications that ensure our users' data is safe and secure" while also mentioning data encryption, a secure cloud architecture, and continuous monitoring; for certification, they write "Grammarly's security controls are validated by enterprise-grade compliances and certifications from external auditors"). However, they barely mention AI legal requirements (not mentioned regarding AI, but they mentioned the privacy part only e.g. GDPR, CCPA). They only say they "comply with privacy regulations and frameworks".

Sources

- https://www.grammarly.com/about
- https://www.grammarly.com/privacy-policy
- https://www.grammarly.com/trust
- https://www.grammarly.com/responsibleai#sectionGroup_6B4OPOYN23X96bQUqtVMxR
- https://www.grammarly.com/acceptable-use-policy

Topic 3:
Responsibility Mindset

3. 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:

3.1 We write **policies and guidelines** about AI ethics.
3.2 We document **roles, responsibilities, and lines of communication** related to AI risk management.
3.3 We provide **training** about AI ethics to relevant personnel.
3.4 We implement practices to foster **critical thinking** about AI risks.

Input Diversity

Mert Cuhadaroglu - Deutsche Telekom, High

Deutsche Telekom has a whistleblower portal named TellMe! which can be used by anybody and anywhere in the World to raise concerns (about anything including violations of human rights). Every concern raised is examined by a team of experts and every whistleblower has the right to be protected from reprisals. https://www.bkms-system.net/bkwebanon/report/clientInfo?cin=dt42017&c=-1&language=eng

All employees, but also business partners, customers, shareholders and other stakeholders who wish to report possible violations of internal Group policies, laws or rules of conduct can submit their messages here. It is stated that absolute confidentiality is guaranteed in all cases.

I have tested the portal, there are two main categories; a) criminal law / breaches of law and policy, b) information / complaints on financial statements and audits. Category a) includes a sub category "the requirements of the code of conduct" and in the code of conduct Deutsche Telekom commits, among other things, to respect and promote human rights. I would have structured it in a way that human rights would be a seperate sub category to choose.

Deutsche Telekom's compliance with the Code of Human Rights & Social Principles is surveyed once a year at all Group companies worldwide as part of the Social Performance Report. I have searched for this report, the last one was for the year 2021 and it included only the outcomes of a survey made with subsidiaries; most of the subsidiaries surveyed did not see any human rights risks. But i could see no results in the report from the whistleblower platform.

Sources:

- https://www.telekom.com
- https://www.cr-report.telekom.com/2022/download-center
- https://www.telekom.com/en/company/data-privacy-and-security/news/transparency-report-363546

Topic 4: Measure Impacts

4. We measure the potential negative impacts of this AI system

4.1 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.
4.2 We have a clear set of **methods and tools** to use when measuring the impacts of this AI system. It includes which metrics and datasets we use.
4.3 We evaluate the **effectiveness of our measurement** processes
4.4 We regularly reevaluate and document the **performance** of this AI system in conditions similar to deployment
4.5 We regularly evaluate **bias and fairness** issues related to this AI system
4.6 We regularly evaluate **privacy** issues related to this AI system
4.7 We regularly evaluate **environmental** impacts related to this AI system
4.8 We regularly evaluate **transparency and accountability** issues related to this AI system
4.9 We regularly evaluate **security and resilience** issues related to this AI system
4.10 We regularly evaluate **explainability** issues related to this AI system
4.11 We regularly evaluate **third-party** issues, such as IP infringement, related to this AI system
4.12 We regularly evaluate **other impacts** related to this AI system
4.13 If evaluations use **human subjects**, they are representative and meet appropriate requirements

Coverage

Katherine Grillaert - Palantir, Medium

Palantir published their "Approach to AI Ethics," a 2200 word communication on their website..In a paragraph discussing monitoring outcomes, Palantir states the need to refine algorithms based on an understanding of unavoidable trade-offs when prioritizing fairness metrics, and calls for documentation of such decisions. This fulfills the requirement for evaluating and documenting bias and fairness for an algorithm, although the frequency of this task for a given deployment is unspecified...

Palantir did publish articles on their official blog titled "Palantir Foundry for AI Governance: Ethical AI in Action"[5] and "Enabling Responsible AI in Palantir Foundry"[6]. These documents mention measurement of bias and fairness, transparency of model training and testing sets, and evaluation of a model's performance. However, they do not meaningfully address the schedule of evaluation, documentation, measurements of risk, or any other standards against which actions should be measured...

Topic 4: Measure Impacts

Coverage

Katherine Grillaert - Cont'd

Palantir's Approach to Ethics document did not give coverage to the categories strategy for measuring the impacts, methods and tools, effectiveness of measurement, and performance. However, their perspective is briefly mentioned in additional sources. In comments regarding the United States government's Federal Engagement in Developing Technical Standards and Related Tools for AI Technologies, Palantir posits that domain- and context-specific metrics should be considered, and that performance be tested in four stages of the model lifecycle, from training to maintenance.[8] In their response to the U.S. Office of Management and Budget Draft Memo for Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence, Palantir calls for the need to have robust testing and evaluation frameworks for large language models in order to identify risks, and to clarify the metrics used for objective assessment.[2] While these comments are a signal of Palantir's thought leadership, there is no further evidence of a commitment to concrete standards.

It is unclear if Palantir had considered the applicability of measuring contextual risk in three categories. These categories are environmental impacts, human subjects, and other impacts. Overall, six substatements were addressed in primary documents (Approach to AI Ethics, Principles), and an additional three substatements were covered in secondary documents (blog, public comments). Four substatements were not covered at all. Primary documents were prioritized when scoring, resulting in a medium score for this section.

Sources:

- Palantir. (n.d.). Palantir AI Ethics. Retrieved February 9, 2024, from https://www.palantir.com/pcl/palantir-ai-ethics/
- Bowman, C., & Jagasia, A. (2023, July 6). RE: Response to the Office of Science and Technology Policy "Request for Information: National Priorities for Artificial Intelligence." Retrieved February 9, 2024, from https://www.palantir.com/assets/xrfr7uokpv1b/6AJzdNE8EeatEI0leaSzKL/8e6c6765ec8b801ee9f36b74986a370f/OSTP_National_Priorities_on_AI.pdf
- Palantir. (n.d.). Principles. Retrieved February 9, 2024, from https://www.palantir.com/pcl/principles/
- Bowman, C. (2023, March 2). The efficacy and ethics of AI must move beyond the performative to the operational. Palantir Blog. Retrieved from https://blog.palantir.com/the-efficacy-and-ethics-of-ai-must-move-beyond-the-performative-to-the-operational-1792e933b34
- Adams, M., & Joshi, I. (2023, April 27). Palantir Foundry for AI Governance: Ethical AI in Action. Palantir Blog. Retrieved from https://blog.palantir.com/palantir-foundry-for-ai-governance-ethical-ai-in-action-d9c6c530beda
- Jagasia, A., McNeal, A., Bowman, C., & Weatherburn, N. (2023, February 8). Enabling Responsible AI in Palantir Foundry. Palantir Blog. Retrieved from https://blog.palantir.com/enabling-responsible-ai-in-palantir-foundry-ac23e3ad7500
- Palantir. (n.d.). Palantir's response to OMB on AI governance, innovation, and risk management. Retrieved February 9, 2024, from https://blog.palantir.com/palantirs-response-to-omb-on-ai-governance-innovation-and-risk-management-1e2be610a6e9
- Palantir Technologies. (2019, July 19). Comment Template for Draft Plan for Federal Engagement in Developing Technical Standards and Related Tools for AI Technologies. Retrieved February 9, 2024, from https://www.nist.gov/system/files/documents/2019/09/16/palantir-technologies-comments-07192019.pdf

Topic 5: Transparency

1. **We document information** about the system, including explaining how it works, limitations, and risk controls

    1. 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.
    2. We document the system **risk controls**, including in third-party components
    3. We **explain the model** to ensure responsible use
    4. We **inventory** information about this AI system in a **repository** of our AI systems

Robustness

Anastasiia Gaidashenko - Meta, Medium

Summary: While Meta documents limitations and provides mechanisms for human oversight, the robustness of these processes appears limited. The documentation addresses basic topics, but the human oversight mechanisms are reactive rather than proactive, indicating a lack of systematic, regular, and comprehensive risk management practices. This is highlighted by the reliance on in-app feedback tools and advisory councils, which are valuable but suggest a more reactive approach to addressing issues.

- Substatement 5.1:
    - Limitations address most basic topics
    - Human oversight mechanisms are rather reactive than proactive
- Substatement 5.2:
    - No Access Controls and Authentication due to open-source nature of the model
- Substatement 5.3:
    - No malicious use preventions
- Substatement 5.4:
    - GitHub allows direct discussion in the repo

Evidence:

- “In-app feedback tools enable people to report responses or image outputs they consider unsafe or harmful. This feedback will be reviewed by humans to determine if our policies have been violated" [2]
- Input and Output Safeguards: "We have also leveraged large language models specifically built for the purpose of helping to catch safety violations" [1]
- Model Reporting: "The research paper and model card provide information about the capabilities and limitations of the models, which will help developers more safely tune, evaluate" [7]

Topic 5: Transparency

Input Diversity

Anastasiia Gaidashenko - Meta, High

Summary: Meta demonstrates high input diversity in documenting system limitations and oversight options. The evidence shows diverse sources of feedback and consultation, including academic researchers, youth and safety advisory councils, and a broad range of internal and external stakeholders involved in red teaming and adversarial testing. This diversity in input sources helps ensure that a wide range of perspectives and concerns are considered in identifying and documenting system limitations and oversight mechanisms.

- Substatement 5.1:
  - The input diversity is notable, with feedback and insights being gathered from a broad spectrum of sources
- Substatement 5.2:
  - Wide range of inputs from various stakeholders, many external partners
- Substatement 5.3:
  - No signs of diverse input here
- Substatement 5.4:
  - Crowdsources info

Evidence:

- "We regularly consult with our Youth Advisory Council and Safety Advisory Council" [2]
- "We're launching a program for academic researchers, [...] where university partners explore topics related to privacy, safety, and security of large language models" [3]
- Red Teaming and Adversarial Testing: "series of red teaming with various groups of internal employees, contract workers, and external vendors [...] included individuals representative of a variety of socioeconomic, gender, ethnicity, and racial demographics" [4]; "we submitted Llama 2 to the DEFCON conference, where it could be stress-tested by more than 2,500 hackers" [5]
- Collaboration with External Partners: "partners are working with us on open trust and safety including: AI Alliance, AMD, Anyscale, AWS, Bain, Cloudflare, Databricks, Dell Technologies, Dropbox, Google Cloud, Hugging Face, IBM, Intel, Microsoft, MLCommons, Nvidia, Oracle, Orange, Scale AI, Together.AI" [6]
- Some information about model is crowdsourced through grants: "Llama Impact Grants"

Topic 5: Transparency

Sources

- 1: Building Generative AI Responsibly https://ai.meta.com/static-resource/building-generative-ai-responsibly/
- 2: Overview of Meta AI safety policies prepared for the UK AI Safety Summit https://transparency.fb.com/en-gb/policies/ai-safety-policies-for-safety-summit/
- 3: Open Innovation AI Research Community https://llama.meta.com/open-innovation-ai-research-community/
- 4: Llama 2: Open Foundation and FineTuned Chat Models https://ai.meta.com/research/publications/llama-2-open-foundation-and-fine-tuned-chat-models/
- 5: On AI, Progress and Vigilance Can Go Hand in Hand https://about.fb.com/news/2024/01/davos-ai-discussions/
- 6: Introducing Purple Llama for Safe and Responsible AI Development https://about.fb.com/news/2023/12/purple-llama-safe-responsible-ai-development/
- 7: Llama 2 Responsible Use Guide https://ai.meta.com/static-resource/responsible-use-guide/
- 8: Responsible Use: How Meta Responds to One of the Central Fears of Open Source AI https://www.newamerica.org/oti/blog/responsible-use-how-meta-responds-to-one-of-the-central-fears-ofopen-source-ai/
- 9: Building Generative AI Features Responsibly https://about.fb.com/news/2023/09/building-generative-ai-features-responsibly/

Topic 6: Risk Mitigation Plan

1. **We plan how to respond to risks**, including setting priorities and documenting residual risks
    1. We **plan** how we will respond to the risks caused by this AI system. The response options can include mitigating, transferring, avoiding, or accepting risks.
    2. 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.
    3. We identify the **residual risks** of this AI system (the risks that we do not mitigate). The documentation includes risks to buyers and users of the system.
    4. We have a plan for addressing **unexpected risks** related to this AI system as they come up

Robustness

Anastasiia Gaidashenko - Meta, Medium

Summary: Meta exhibits high engagement in risk mitigation through initiatives like red teaming, model fine-tuning for safety, and the development of the Open Loop program, indicating proactive measures. However, the lack of detailed mechanisms for prioritizing responses based on impact, likelihood, and other factors, combined with an absence of concrete plans for addressing unexpected risks, suggests that while robust mechanisms are in place, there's room for improvement in depth and systematic application. Therefore, a medium score for robustness reflects proactive but not fully systematic risk management practices.

- Substatement 6.1: Use of red teaming, fine-tuning models for safety, and the development of the Open Loop program
- Substatement 6.2: No data
- Substatement 6.3: Pro: commitment to continuous improvement and collaboration. Con: shallow analysis
- Substatement 6.4: depth of the response mechanism is not fully detailed. Evidence:
    - "Open Loop program is launching its first policy prototyping program in the United States, which is focused on the National Institute of Standards and Technology (NIST) AI Risk Management Framework (AI RMF) 1.0" [2]
    - Purple Llama project aims to provide tools to help developers assess and improve trust and safety in AI models, thereby mitigating potential risks.
    - The licensing agreement for Llama 2 includes restrictions to prevent activities that present a risk of death or bodily harm.
    - "Dedicated teams of experts have spent thousands of hours stress-testing these models, looking for unexpected ways they might be used along with identifying and fixing vulnerabilities

Topic 6:
Risk Mitigation Plan

Robustness

Anastasiia Gaidashenko - Meta, Medium, Cont'd

Sources:

- 1: Building Generative AI Responsibly https://ai.meta.com/static-resource/building-generative-ai-responsibly/
- 2: Overview of Meta AI safety policies prepared for the UK AI Safety Summit https://transparency.fb.com/en-gb/policies/ai-safety-policies-for-safety-summit/
- 3: Open Innovation AI Research Community https://llama.meta.com/open-innovation-ai-research-community/
- 4: Llama 2: Open Foundation and Fine-Tuned Chat Models https://ai.meta.com/research/publications/llama-2-open-foundation-and-fine-tuned-chat-models/
- 5: On AI, Progress and Vigilance Can Go Hand in Hand https://about.fb.com/news/2024/01/davos-ai-discussions/
- 6: Introducing Purple Llama for Safe and Responsible AI Development https://about.fb.com/news/2023/12/purple-llama-safe-responsible-ai-development/
- 7: Llama 2 Responsible Use Guide https://ai.meta.com/static-resource/responsible-use-guide/
- 8: Responsible Use: How Meta Responds to One of the Central Fears of Open Source AI https://www.newamerica.org/oti/blog/responsible-use-how-meta-responds-to-one-of-the-central-fears-of-open-source-ai/
- 9: Building Generative AI Features Responsibly https://about.fb.com/news/2023/09/building-generative-ai-features-responsibly/
- 10: My spreadsheet with analysis results https://docs.google.com/spreadsheets/d/1Is7vKo-QA4txTyFL7blLmDHdeqMO6bnpp7BUzVbeRt0/edit?usp=sharing

Topic 7:
Risk Mitigation Activities

1. **We act to minimize risks**, including addressing our prioritized risks and tracking incidents:
    1. We proactively evaluate **whether this system meets its stated objectives** and whether its development or deployment should proceed
    2. We ensure this AI's **bias and fairness** performance meets our standards
    3. We ensure this AI's **privacy** performance meets our standards
    4. We ensure this AI's **environmental** performance meets our standards
    5. We ensure this AI's **transparency and accountability** meets our standards
    6. We ensure this AI's **security and resilience** meets our standards
    7. We ensure this AI's **explainability** performance meets our standards
    8. We ensure this AI's **third-party** impacts, such as **IP infringement**, meet our standards
    9. We implement processes for **human oversight** related to this AI system
    10. We implement processes for **appeal** related to this AI system
    11. 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.
    12. We address **all other risks prioritized in our plans** related to this system by conducting measurable activities
    13. We address **unexpected risks** related to this system by conducting measurable activities
    14. We track and respond to **errors and incidents** related to this system by conducting measurable activities

Coverage

Anonymous - Khan Academy, Medium

Positives:

- They have published a set of AI guidelines that shares their approach to building AI responsibly (https://blog.khanacademy.org/aiguidelines/?)
- They attest to have "studied and adapted frameworks from the National Institute of Standards and Technology (NIST) and the Institute for Ethical AI in Education to evaluate and mitigate AI risks specific to Khan Academy." which is commendable
- They conduct fine-tuning, prompt engineering and 'red teaming' to mitigate risk
- They seem to have a good moderation framework to flag inappropriate interactions with the AI, and also adopt an Adult-In-The-Loop framework, sharing student chats with teachers and parents
- They seem to be launching this tool in a phased approach, which is good, as I assume they are looking to launch it correctly: "We limit access to our AI through Khan Labs, a space for testing learning tools. We use careful selection criteria so that we can test features in Khan Labs before broadening access."

Topic 7: Risk Mitigation Activities

Coverage

Anonymous - Cont'd

Continuation of Positives:

- They provide a set of AI literacy tools and an AI for Education course, which contains useful information, however, seems to take a very positive approach to introducing AI and does not focus enough on risks and mitigation of those risks.
- They seem to cover: privacy performance, some transparency, have some human oversight/moderation practices, a process for appeal if the account got deactivated, and the ability to give feedback.

Negatives:

- There are many substatements that they do not cover in their public documentation. The most glaring are around metrics about the performance of the system, whether this system meets its stated objectives, whether bias and fairness performance meets their standards, no mention of environmental performance - They mention: "Individuals and teams are asked to identify ethical considerations and evaluate risks at the outset of every project. Our decision making is guided by risk evaluation. We prioritize risk mitigation, we embrace transparency, and we continuously reflect on the impact of our work.
- We have a detailed monitoring and evaluation plan in place during this testing period. We will learn, iterate, and improve." However, not much information is provided about the monitoring and evaluation plan. Given this is an educational platform with such a high impact and high reach, I would expect much more information about how they ensure their systems are complying with the standards they set and with the substatements above.
- They seem to rely on the users (parents, teachers, students) for providing feedback via a feedback form embedded in the chatbot. However, given this is an education tool, I would expect that assessment of the system does not solely fall on the shoulders of the users
- In terms of transparency and explainability, they have a few statements pointing out that the AI can make mistakes and give some advice about how to interact with LLMs (https://support.khanacademy.org/hc/en-us/articles/13888935335309-How-do-the-Large-LanguageModels-powering-Khanmigo-work). However, given the educational context, I find that lacking. For example, they have activities such as chatting with a historical figure, with seemingly no explainability attached to the responses or disclaimers. Given what we know about bias in datasets and the lack of diverse historical perspectives, I find it difficult to believe, without seeing additional public documentation from them, that they are doing specific work to address bias such as Western-centric historical perspectives, white saviorism etc.
- Quite concerning, they do not seem to have enough information or educational materials about sharing personal information. They mention: "Remember: Khanmigo Lite GPTs are an AI tool. We recommend that you do NOT share personal data when using Khanmigo Lite or any other large language model (LLM)." and the chatbot input box contains text says: "Type message (do NOT share any personal data)", but not much else is in place here. Given this tool is used by children, I would expect more. Their privacy notice (https://support.khanacademy.org/hc/en-us/articles/22396485532173-Khanmigo-Lite-Privacy-Notice) mentions: "When you use Khanmigo Lite, use of your data is subject to OpenAI's privacy policy and controls and usage policies. Please refer to OpenAI's privacy policy for information about your privacy choices when using GPTs, including your choices regarding whether OpenAI can use your chats to train its models.", which implies that this is left to the latitude of OpenAI's policies. However, I am unsure where the child data protection acts they comply with (Children's Online Privacy and Protection Act (COPPA), the Family Educational Rights and Privacy Act (FERPA), and the Student Online Personal Information Protection Act (SOPIPA)) fall in this arragement.

Topic 8:
Pre-deployment Checks

1. **We only release versions that meet our AI ethics standards**
    1. We demonstrate that this system is **valid, reliable**, and meets our standards. We document the conditions under which it falls short.

Coverage

Caroline Lancelot - Grammarly, Medium

- They claim the system is valid and reliable without clearly demonstrating it.
- They mention their security infrastructure is built upon industry standards.
- They also mention Enterprise-grade attestations validate our security controls (See the attestations and certifications that ensure our users' data is safe and secure) and say they are Trusted by Thousands of Organizations Around the World.
- They do not clearly document the condition under which it falls short.
- This deserves a low to medium grade.

Robustness

Caroline Lancelot - Grammarly, Low

They do not talk about the Regularity; Systematicity; Trained Personnel; Sufficient Resources; Adaptivity; Cross-functionality of pre-deployment checks.

Input Diversity

Caroline Lancelot - Grammarly, Medium

- They do not mention if and how external participants or organizations are used for pre-deployment checks.
- They always mention using user feedback without being clear at which phase they are used exactly

Documents used throughout

- https://www.grammarly.com/about
- https://www.grammarly.com/privacy-policy
- https://www.grammarly.com/trust
- https://www.grammarly.com/responsibleai#sectionGroup_6B40P0YN23X96bQUqtVMxR
- https://www.grammarly.com/acceptable-use-polic

Topic 9: Monitoring

1. **We monitor and resolve issues as they arise**
    1. We **plan how to monitor** risks related to this system post-deployment
    2. We monitor this system's **functionality and behavior** post-deployment
    3. We apply mechanisms to **sustain the value of this AI system** post-deployment
    4. We capture and evaluate **input from users** about this system post-deployment
    5. We monitor **appeal and override** processes related to this system post-deployment
    6. We monitor **incidents** related to this system **and responses** to them post-deployment
    7. We monitor incidents related to **high-risk third-party** components and respond to them
    8. We implement **all other** components of our post-deployment monitoring plan for this system
    9. We monitor issues that would trigger our **end-of-life mechanisms** for this system, and we take the system offline if issues come up

Coverage

Andrew McAdams - Inworld, Medium

Inworld scores fairly well here. Their blog post on safety specifically mentions ongoing monitoring and improvement of the system(s) in section 5 (sub-statement 9.1). Sections 3-5 on the blog post detail developer controls, reporting and moderation functions, and the ongoing monitoring and improvement of their systems. (sub-statement 9.2).

These same systems suffice to maintain the value of this AI system for their users, post-deployment (sub-statement 9.3), section 4 details a reporting and feedback mechanism within the product, allowing for input from users about the performance of the systems and adherence to the policies and code of conduct (sub-statement 9.4).

I cannot determine from external documentation whether sub-statement 9.5 applies here.

Section 5 of the blog post satisfies sub-statement (9.6) for monitoring of incidents and responses.

Sub-statement 9.7 is difficult to gauge in light of the lack of explicit material. I attempted to rely on other certifications that would have similar requirements, such as SOC2 Type 2 certification or ISO27001, ISO42001 but Inworld makes no references to any of those certifications in their external documentation. Their Terms of Service do not reference monitoring their third-party components for incidents and claim no liability for those third-party components. In the absence of evidence supporting this, I must conclude that Inworld does not satisfy this requirement.

Topic 9: Monitoring

Coverage

Andrew McAdams - Cont'd

While not explicitly called out, Sub-statement H is likely satisfied through the other monitoring commitments made in the blog post.

Sub-statement I is not referenced, so Inworld fails to meet these requirements.

As it stands, there's enough externally available information about the monitoring of the system for me to feel confident about those internal processes. I would rate Inworld's score for coverage of monitoring at Acceptable / Medium. To increase this score I would expect to see more transparency about the systems themselves to better establish overall coverage. Additionally, related certifications that have controls for many of these things, like SOC2 Type 2 certification, ISO27001, and ISO42001 would demonstrate satisfactory compliance with these sub-statements without requiring additional disclosure.

Robustness

Andrew McAdams - Inworld, Medium

[T]he monitoring topic suffers from many of the same issues when reviewing only publicly available information. It's difficult to rate the criteria as much of this isn't disclosed publicly. However, it's possible to glean more about monitoring from the available documentation.

The existence of a monitoring section (section 5) in the Safety blogpost implies both regularity of review and adaptability. While the monitoring function is going to be primarily support and R&D, we can also assume that there's training across these functions to help support the monitoring.

I cannot make any assumptions about the sufficient resourcing and cross-functional buy-in for the same reason as before - the information isn't available, either explicitly or implied. Lacking any indication of meeting the requirement here, I assume that Inworld fails to meet the criteria in these areas.

I think they have achieved a low-to-medium robustness score because while there are a lot of indicators that they've satisfied the criteria for the score, I'm making a lot of assumptions and reading between the lines.

To improve this score, Inworld should be more transparent about the plans, responses, and leadership buy-in. In many cases, a single disclosure will likely satisfy many of the topic ratings and improve the ratings across the board (assuming the disclosure is sufficiently detailed). Additionally, as before other, ancillary certifications will help establish that Inworld meets some criteria without requiring additional disclosure or attestation.

Sources (throughout):
- https://inworld.ai/blog/inworlds-commitment-to-safety
- https://docs.inworld.ai/docs/resources/safety/
- https://inworld.ai/terms

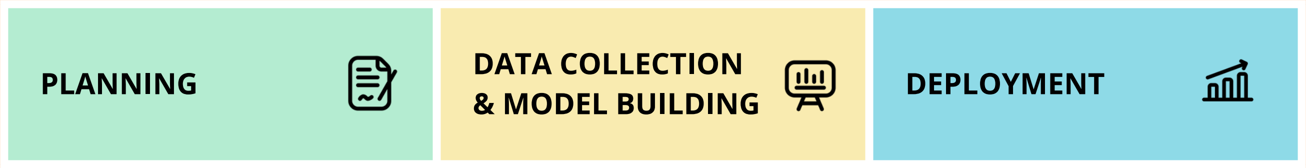
Appendix:

The Full Questionnaire & Scoring Guidelines

Responsible AI Governance Maturity Model
Full Questionnaire & Scoring Guidelines

Based on the NIST AI RMF   Full paper: Dotan et al. (2024)
Supported by the Notre Dame-IBM Tech Ethics Lab

At a Glance - Across 3 Development Phases



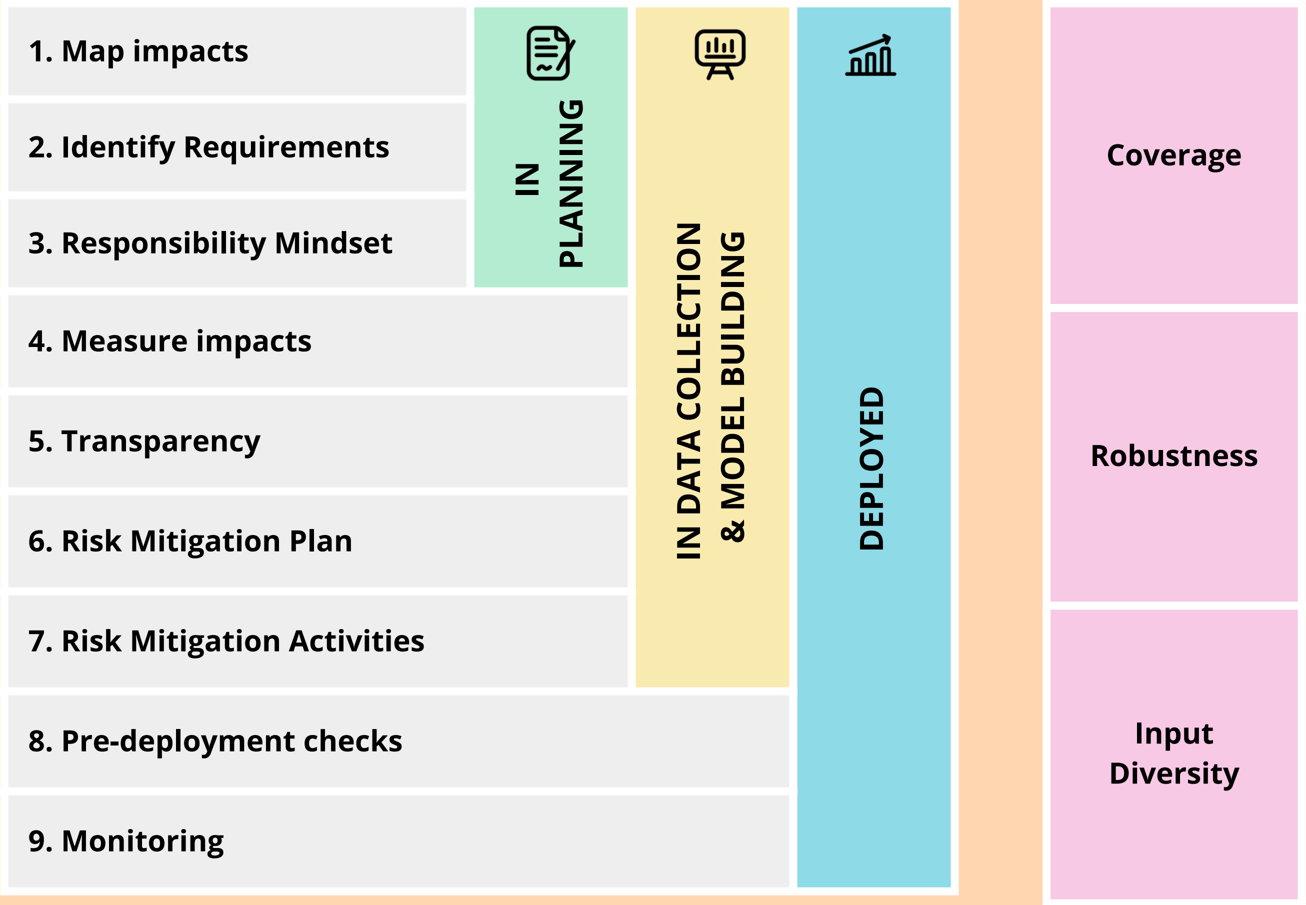
PLANNING

DATA COLLECTION & MODEL BUILDING

DEPLOYMENT

Questions - 9 Topics Across the 3 Phases

Metrics



1. Map impacts
2. Identify Requirements
3. Responsibility Mindset
4. Measure impacts
5. Transparency
6. Risk Mitigation Plan
7. Risk Mitigation Activities
8. Pre-deployment checks
9. Monitoring

IN PLANNING

IN DATA COLLECTION & MODEL BUILDING

DEPLOYED

Coverage

Robustness

Input Diversity

Scoring Guidelines: Metrics

The score of each topic should be based on the three metrics below. The evaluator ranks how well each metric is satisfied: Low, medium, or high.

Coverage	The scoring of each topic should be higher the better the coverage of the activities in the substatements.
Robustness	Scores should be higher the more the activities are robust. Activities that are robust share the following characteristics: • **Regularity** - Performed in a routine manner • **Systematicity** - Follow policies that are well-defined and span company-wide • **Trained Personnel** - Performed by people who are properly trained and whose roles in the activities are clearly defined • **Sufficient Resources** - Supported by sufficient resources, including budget, time, compute power, and cutting-edge tools • **Adaptivity** - Adapting to changes in the landscape and product, including regular reviews and effective contingency processes to respond to failure • **Cross-functionality** - All core business units and senior management 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 the activities are informed by 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, such as members of one internal team only. • **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".

Scoring Guidelines: Explanations

Scores must be accompanied with an explanation. The explanation should refer to information about what the organization does or doesn't do and any relevant contextual facts.

Resources - Where to find relevant information? (Examples)

Internal information (if available)	• Internal documents • Interviews with employees • Informal conversations and employee knowledge • Internal metrics, e.g., Objectives and Key Results (OKRs) and Key Performance Indicators (KPIs)
External information	• External company documents, e.g. AI ethics frameworks • External company reports, e.g., annual or ESG reports • Research papers by or about the company • Media reports • Lawsuits

Evaluation Elements - What to evaluate when scoring? (Examples)

Execution	Outcomes, procedures, and resources dedicated to activities: ○ RAI metrics and progress on those metrics ○ How much time is spent working on certain tasks ○ The execution of RAI best practices, such as red-team exercises or ethics reviews
Uptake	How relevant activities and deliverables are received: ○ Whether the outputs of RAI work are officially adopted by the company ○ Whether and how the company's leadership supports it
People	Who performs the relevant activities and how: ○ Whether the people conducting the relevant activities are doing so as part of their official capacity or as a voluntary side project ○ The number of people assigned to relevant tasks ○ How suitable these people are to perform the tasks
Communication	The nature of internal conversations related to the relevant activities: ○ How frequently relevant topics or tasks are discussed ○ The depth of the conversations ○ The formality of conversation channels for relevant topics. For example, is there a dedicated time to talk about it, or does it come up occasionally in Slack?

Scoring Guidelines: Overall Calculation

The overall score of each topic is calculated based on the scores of all the metrics. It ranges from 1-5, where 1 is the lowest and 5 is the highest.

Score	Conditions	Shorthand
5	All three metrics are satisfied to a high degree	HHH
4	Two of the metrics are satisfied to a high degree and one to a medium degree	HHM
3	One of the following is the case: • Two of the metrics are satisfied to a medium degree and one to a high degree • Two of the metrics are satisfied to a high degree and one to a low degree • One metric is satisfied to a high degree, one to a medium degree, and one to a low degree. • All three metrics are satisfied to a medium degree.	HMM HHL HML MMM
2	One of the following is the case: • Two of the metrics are satisfied to a medium degree and one to a low degree • One metric is satisfied to a medium degree and two to a low degree • One of the metrics is satisfied to a high degree and two to a low degree.	MML MLL HLL
1	All metrics are satisfied to a low degree	LLL

Full Questionnaire: All Phases

1. Planning: Map Impacts We clearly define what the AI is supposed to do and its impacts, including scope, goals, methods, and negative and positive potential impacts of these activities.		
1.1	**Goals**	We define the goals, scope, and methods of this AI system.
1.2	**Positive Impacts**	We identify the benefits and potential positive impacts of this AI system, including the likelihood and magnitude.
1.3	**Business Value**	We identify the business value of this AI system.
1.4	**Negative Impacts**	We identify the possible negative impacts of this AI system, including the likelihood and magnitude.
1.5	**Costs of Malfunction**	We identify the potential costs of malfunctions of this AI system, including non-monetary costs such as decreased trustworthiness.
1.6	**Unexpected Impacts**	We implement processes to integrate input about unexpected impacts.
1.7	**Methods and Tools**	We identify the methods and tools we use for mapping impacts.

Full Questionnaire: All Phases

2. Planning: Identify Requirements

We identify the requirements the AI must meet, including compliance, certifications, and human oversight needs.

2.1	**Human Oversight**	We identify the human oversight processes the system needs.
2.2	**Certifications**	We identify the technical standards and certifications the system will need to satisfy.
2.3	**Legal Requirements**	We identify AI legal requirements that apply to this AI system.

3. Planning: Responsibility Mindset

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.

3.1	**Policies and Guidelines**	We write policies and guidelines about AI ethics.
3.2	**Roles and Responsibilities**	We document roles, responsibilities, and lines of communication related to AI risk management
3.3	**Training**	We provide training about AI ethics to relevant personnel.
3.4	**Critical Thinking**	We implement practices to foster critical thinking about AI risks.

Full Questionnaire: Data Collection & Model Building + Deployment Phases

4. Data Collection & Model Building: Measure Impacts

We measure potential negative impacts.

4.1	**Strategy for Measuring the Impacts**	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.
4.2	**Methods and Tools**	We have a clear set of methods and tools to use when measuring the impacts of this AI system. It includes which metrics and datasets we use.
4.3	**Effectiveness**	We evaluate the effectiveness of our measurement processes.
4.4	**Performance**	We regularly revaluate and document the performance of this AI system in conditions similar to deployment.
4.5	**Bias and 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.
4.7	**Environmental**	We regularly evaluate environmental impacts related to this AI system.
4.8	**Transparency and Accountability**	We regularly evaluate transparency and accountability issues related to this AI system.
4.9	**Security and Resilience**	We regularly evaluate security and resilience issues related to this AI system
4.10	**Explainability**	We regularly evaluate explainability issues related to this AI system
4.11	**Third-party**	We regularly evaluate third-party issues, such as IP infringement, related to this AI system.
4.12	**Other Impacts**	We regularly evaluate other impacts related to this AI system.
4.13	**Human Subjects**	If evaluations use human subjects, they are representative and meet appropriate requirements.

Full Questionnaire: Data Collection & Model Building + Deployment Phases

5. Data Collection & Model Building: Transparency

We document information about the system, including explaining how it works, limitations, and risk controls.

5.1	**Human 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.
5.2	**Risk Controls**	We document the system risk controls, including in third-party components.
5.3	**Model Explanation**	We explain the model to ensure responsible use.
5.4	**Inventory**	We inventory information about this AI system in a repository of our AI system.

6. Data Collection & Model Building: Risk Mitigation Plan

We plan how to respond to risks, including setting priorities and documenting residual risks.

6.1	**Plan**	We plan how we will respond to the risks caused by this AI system. The response options can include mitigating, transferring, avoiding, or accepting risks.
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.
6.3	**Residual Risks**	We identify the residual risks of this AI system (the risks that we do not mitigate). The documentation includes risks to buyers and users of the system.
6.4	**Unexpected Risks**	We have a plan for addressing unexpected risks related to this AI system as they come up.

Full Questionnaire: Data Collection & Model Building + Deployment Phases

7. Data Collection & Model Building: Risk Mitigation Activities

We act to minimize risks, including addressing your prioritized risks and tracking incidents.

7.1	**Meets Objectives**	We proactively evaluate whether this system meets its stated objectives and whether its development or deployment should proceed.
7.2	**Bias and Fairness**	We ensure this AI's bias and fairness performance meets our standards.
7.3	**Privacy**	We ensure this AI's privacy performance meets our standards.
7.4	**Environmental**	We ensure this AI's environmental performance meets our standards.
7.5	**Transparency and Accountability**	We ensure this AI's transparency and accountability meets our standards.
7.6	**Security and Resilience**	We ensure this AI's security and resilience meets our standards,
7.7	**Explainability**	We ensure this AI's explainability performance meets our standards.
7.8	**Third-party**	We ensure this AI's third-party impacts, such as IP infringement, meet our standards.
7.9	**Human Oversight**	We implement processes for human oversight related to this AI system.
7.10	**Appeal**	We implement processes for appeal related to this AI system.
7.11	**End-of-life Mechanisms**	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.
7.12	**All Other Risks**	We address all other risks prioritized in our plans related to this system by conducting measurable activities.
7.13	**Unexpected Risks**	We address unexpected risks related to this system by conducting measurable activities.
7.14	**Errors and Incidents**	We track and respond to errors and incidents related to this system by conducting measurable activities.

Full Questionnaire: Deployment Phase

8. Deployment: Pre-Deployment Checks We only release versions that meet our AI ethics standards.		
8.1	**Valid and Reliable**	We demonstrate that this system is valid, reliable, and meets our standards. We document the conditions under which it falls short.

9. Deployment: Monitoring We monitor and resolve issues as they arise.		
9.1	**Monitoring Plan**	We plan how to monitor risks related to this system post-deployment.
9.2	**Functionality and Behavior**	We monitor this system's functionality and behavior post-deployment.
9.3	**Sustain Value**	We apply mechanisms to sustain the value of this AI system post-deployment.
9.4	**Input from Users**	We capture and evaluate input from users about this system post-deployment.
9.5	**Appeal and Override**	We monitor appeal and override processes related to this system post-deployment.
9.6	**Incidents and Response**	We monitor incidents related to this system and responses to them post-deployment.
9.7	**High-risk Third-party**	We monitor incidents related to high-risk third-party components and respond to them.
9.8	**All Other Components**	We implement all other components of our post-deployment monitoring plan for this system.
9.8	**End-of-life Mechanisms**	We monitor issues that would trigger our end-of-life mechanisms for this system, and we take the system offline if issues come up.