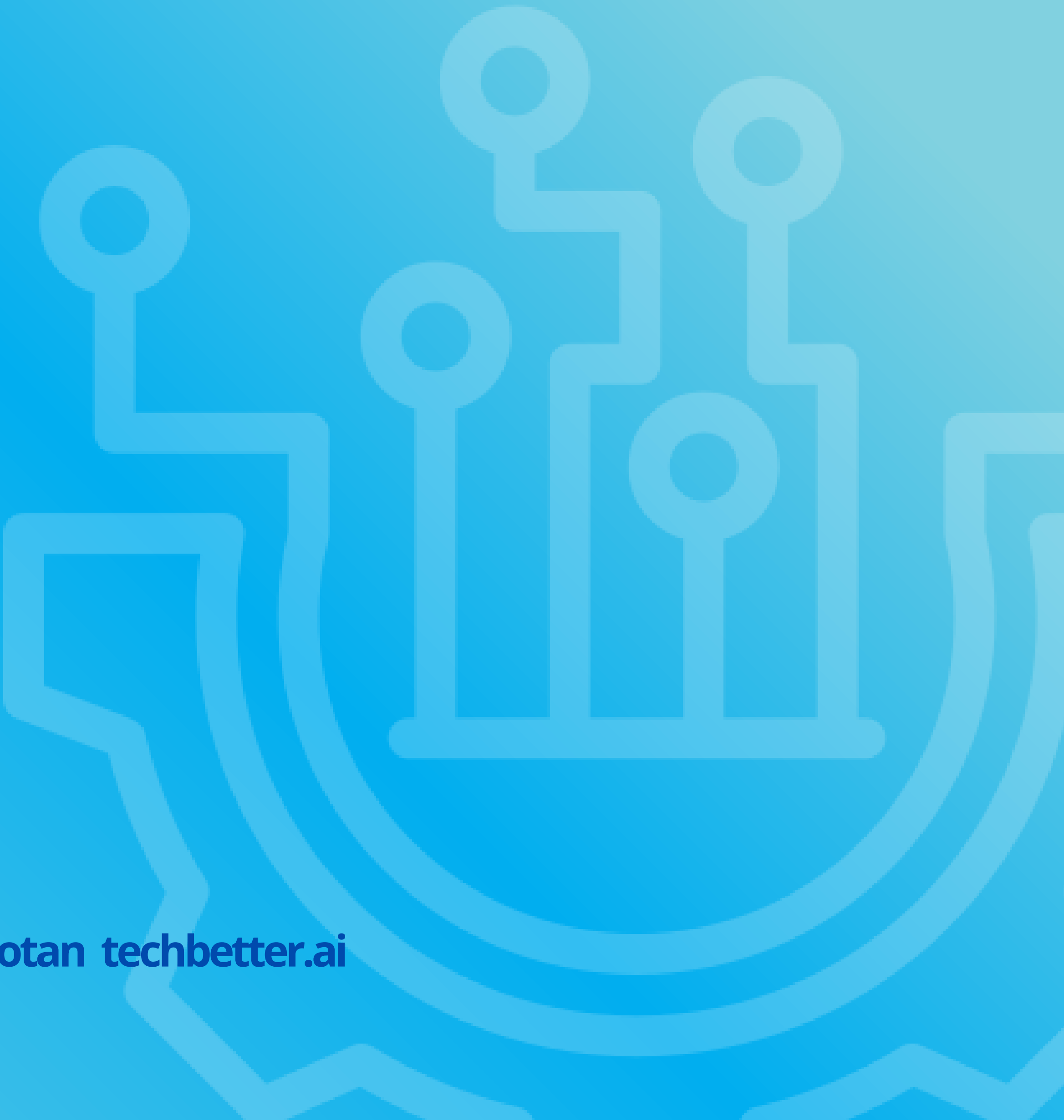
Evaluating & Improving
AI Governance

TechBetter

Early-Stage AI Governance

Framework & Case Studies



Ravit Dotan techbetter.ai

Foreword

Figuring out how to govern AI responsibly, maximize the benefits, and minimize the risks, is a barrier for many organizations. For example, in a recent survey, 52% of executives said that they actively discourage generative AI adoption. The lack of a responsible AI strategy was the second most common reason.

These concerns are appropriate because the stakes for organizations are high, including risks and rewards to the quality of their products, reputation, client attraction, employee attraction, and compliance. The stakes for end users, related communities, and society at large are also high, including mass disinformation, discrimination, privacy violations, and physical and psychological harm, to name just a few.

This report presents a framework, case studies, and insights about evaluating and improving AI governance in companies that develop or deploy it. It focuses on the earliest stage of the development lifecycle, the ideation phase, demonstrating how AI responsibility is crucial even at this stage.

Readers are provided with the following:

- A method for evaluating and improving AI responsibility, building on Dotan et al. (2024) and the NIST AI Risk Management Framework.
- Detailed examples of governance evaluations from early-stage projects that used the framework. The evaluations were a part of a competition at the AI4Gov masters program, and they include risk triage, evaluating current governance activities, and improvement plans.
- Insights arising from the competition, including top prioritized risks, common governance strengths and weaknesses, and strategic plans for early stage projects

For more and to get in touch

For more information about the framework and additional case studies: https://www.techbetter.ai/rai-maturity-model

For support in implementing this process in your organization and for hosting similar competitions at your event, get in touch here: https://www.techbetter.ai/contact

Ravit Dotan,TechBetter

www.techbetter.ai

/ravit-dotan

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International

Table of **Contents**

1. Introduction

2. High Level: Insights and Case Study Summaries

3. Deep Dive: Risk Triaging

4. Deep Dive: Governance Evaluation & Strategy

1. Introduction

AI responsibility, the framework, and the competition

About
This Document

In this report

This report presents a framework for responsible AI governance in organizations that develop or deploy AI, as well as detailed case studies and insights from implementation.

- **The framework** builds on the Responsible AI Governance Maturity Model developed by Dotan et al. (2024), which is based on the NIST AI Risk Management Framework (RMF), a leading AI governance framework. The full framework and additional materials are available at the following link: https://www.techbetter.ai/rai-maturity-model

- **The case studies** are the evaluations carried out by the winners of a competition that took place as part of the AI4Gov masters program. In the competition, groups applied the framework to AI-enabled products at the ideation phase, illustrating how responsible AI governance is crucial even at this very early stage. All products aimed to serve pubic-sector tourism functions.

- **The insights** stem from analyzing the evaluations of all the groups in the competition: Highlighting the importance of AI Responsibility in early stages, top ranked risks, average governance scores, top governance strengths and weaknesses, and top priorities for governance improvement.

Section overview

The structure of this report is as follows:

- **Section 1: Introduction** - Including an overview of the framework and competition.
- **Section 2: High-level insights and case studies** - Including analysis and summaries of the winners' evaluations
- **Section 3: Deep Dive into Risk Triage** - Including detailed examples from the winners' evaluations
- **Section 4: Deep Dive into Governance Evaluation and Strategy** - Including detailed examples from the winners' evaluations

About
AI Responsibility

Responsible AI Governance

Governing AI responsibly means developing and deploying AI in ways that minimize risks and maximize the benefits to people, society, the planet, and the business. The process includes:

- **Risk priorities** - Triaging the risks the AI system poses to all relevant stakeholders.
- **Governance evaluation** - Evaluating the current AI risk management efforts.
- **Governance growth strategy** - Planning how to improve AI governance.
- **Growth strategy implementation** - Implementing the strategy.

The Benefits of AI Responsibility

Responsible AI governance benefits all stakeholders. For end users, impacted communities, and other society in general, AI responsibility increases the likelihood of benefiting from the AI's intended use and decreases the risk of discrimination, privacy violations, physical and psychological harm, financial exploitation, and other harms.

For organizations that develop and deploy AI, the impacts of AI responsibility include the following:

- **Product quality** - While AI can be very effective, it often produces low-quality outputs, such as false information, discriminatory recommendations, analysis errors, and distorted images. AI responsibility improves outcomes and decreases this risk.
- **Reputation** - Poor AI outcomes can tarnish the reputation of everyone involved, including both AI developers and deployers.
- **Compliance** - Irresponsible development and use of AI tools may infringe on AI-specific laws, such as the EU AI Act, and general laws, such as non-discrimination, consumer protection, privacy, and copyright laws.
- **Client attraction and retention** - Surveys show that individual and business consumers are likely to refuse to work with vendors if concerns about to AI irresponsibility arose.
- **Talent attraction and retention** - Surveys show that employees prefer workplaces with ethical leadership.

About
The Evaluation Process

Risk Triaging Questionnaire

The first step is prioritizing the product's AI risks. The process includes going over a list of prominent risks (based on the NIST AI RMF) and ranking each risk from low to high priority based on the answers to two questions:

- **Stakeholder impacts -** How could the use of this AI harm your target audience or other stakeholders?
- **Company impacts -** How could the use of this AI harm your own organization?

Evaluating Governance

The second step is evaluating the organization's current AI risk management efforts by scoring statements about the organization's current governance activities, from 0, which stands for "not at all performing the activities," to 4, which stands for "excellent performance." The process uses the questionnaire and scoring guidelines developed by Dotan et al. (2024), which is based on the NIST AI RMF. The statements in the questionnaire are divided into nine topics across three phases of the development life-cycle (planning, development, and deployment).

Establishing Growth Strategy

The third step is to determine how the organization will improve its AI risk management. To do so, organizations go over the same list of statements about their current activities and rank their priorities: Immediate priority, short term priority, long term priority, or not a priority.

Evaluation Outcomes and Implementation

At the end of the process, organizations have an actionable plan for improving their AI governance based on their context and and an industry standard (the NIST AI RMF). In addition, they have documentation they can leverage with customers, employees, thought leadership, and marketers.

Framework Overview: Responsible AI Governance Maturity Model & Strategy

At a Glance

IDENTIFY YOUR AI RISKS

EVALUATE YOUR GOVERNANCE

DETERMINE NEXT STEPS

Key Risks

- Performance
- Safety
- Privacy
- Security
- 3rd-party / IP
- Fairness
- Ecology
- Explainability
- Transparency

Governance Topics

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

PLANNING (1–3)

DEVELOPMENT (1–7)

DEPLOYMENT (1–9)

Priorities

- Immediate
- Short term
- Long term
- Not a priority

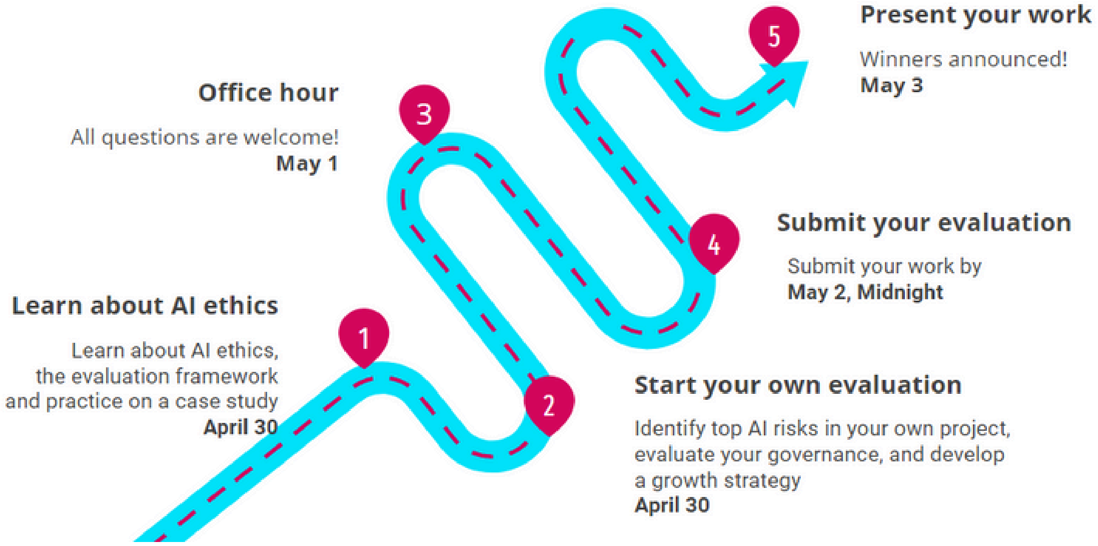
About
The Competition

The Host: AI4Gov Master‘s Program

The competition took place as part of the AI4Gov Masters Program, which provides training for designing, implementing, and governing AI projects in the public sector. During the program, groups ideate AI-enabled tools for the public sector. In the 2024 cohort, the task was to ideate tools to assist public sector tourism agencies. The competition took place at late stages of the ideation process.

Competition Components and Roadmap

The competition included two virtual webinars and one office hour in between, in the span of a week. In the first webinar (3h), the groups were educated about AI ethics and the framework, practiced on a case study, and started evaluating their own projects. After the webinar, the groups continued to evaluate their projects independently and received support during the office hour (45 min). The groups presented their work in the second webinar (1.5h) and then the winners were announced.



Learn about AI ethics
Learn about AI ethics, the evaluation framework and practice on a case study
April 30

Start your own evaluation
Identify top AI risks in your own project, evaluate your governance, and develop a growth strategy
April 30

Office hour
All questions are welcome!
May 1

Submit your evaluation
Submit your work by
May 2, Midnight

Present your work
Winners announced!
May 3

Competition Tasks

The competition included three tasks:

- **Top risks** - Each group discussed a list of nine risks and ranked them from low to high priority. (See section 3 for examples.)

- **Governance strengths** - Each group evaluated its current AI governance efforts using the framework described above. Since all the projects were in the planning stage, they only used the first three topics, which pertain to the planning stage: Map impacts, identify requirements, and AI ethics mindset. Each group ranked itself on all the statements in these topics and provided an explanation for each score. (See section 4 for examples.)

- **The groups' governance improvement priorities** - Each group planned how to improve their AI governance by assigning priorities to the statements relevant for the planning stage, from immediate priority to long term priority. (See section 4 for examples.)

Participants

The competition included 6 groups and 44 participants overall.

All the participants were masters students with strong backgrounds in public-sector work: Some work in the public sector (65.9%) and others work with it (34.1%). They were affiliated with public administration agencies, civil society organizations, tech companies, and other organizations.

Participants had varying levels of AI, design, and ethics competence, from none to very good.

Demographically, the group included participants of 20 different nationalities, coming from 4 different continents, and living in 25 different countries. The average age was 39.8, and the gender distribution was as follows: 14 identify as female, 28 as male, and 2 preferred not to disclose their gender.

2.
High-level: Insights and Case Studies

Including high-level reflections and summaries of winners' evaluations

Insights
Early-Stage Responsibility

The Value of AI Responsibility in Early-Stage

The participants' feedback demonstrates the importance of evaluating and strategizing about AI responsibility even at the planning phase, the very beginning of the development life-cycle. The groups expressed that the process was helpful and important. Some even expressed that they wished they would have gone through it earlier in the ideation phase.

> *"It was a very valuable experience, and it made us realize we haven't thoroughly reflected on risk factors in our project. We now started this reflection and we are convinced it will make our project more comprehensive...it would be much more beneficial if this module came earlier in the process."*
>
> *-- Group 6*

> *"[We learned the] importance of incorporating security considerations from the initial design phase through to post-implementation monitoring and evaluation... [and the] importance of having dedicated professionals—such as security experts, technical architects, and legal advisors—involved at each stage to address specific vulnerabilities and compliance requirements... engaging with users directly enhances the effectiveness of security measures, providing valuable feedback that can be used to fine-tune the app and ensure it meets user expectations in terms of both functionality and safety."*
>
> *-- Group 4*

> *"We found the questionnaire especially helpful in planning and prioritising next steps. We have considered quite a wide range of aspects when designing the solution, but we haven't thought enough about how to execute different steps, in what order, etc."*
>
> *-- Group 3*

Insights
Top Risks & Scores

Top Risks

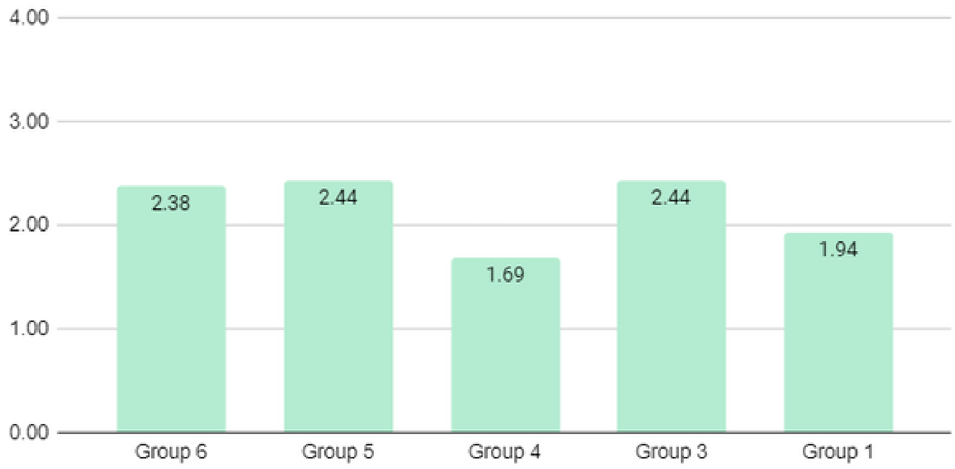
All the groups planned products for the same sub-sector, public administration tourism-related agencies, using similar technologies, LLMs and generative AI. Overall, the risks the groups ranked the highest were performance, privacy, fairness, and security.

Top Risks	
Performance	**Fairness**
Privacy	**Security**

Average Governance Scores

As you may recall, groups ranked statements about their current AI governance activities on a scale of 0-4, where 0 means "not at all performing the activities" and 4 means "excellent performance". The averages scores for each group range between 1.69 - 2.44, with an overall average of 2.18. (Group 2 was excluded because they didn't complete their evaluation.)

Average governance scores per group



Insights
Strengths & Weaknesses

Top Governance Strengths

On average, the groups scored themselves highest on identifying what the AI system is supposed to do and its positive impacts, which is to be expected at the ideation stage. Groups also reported strength in receiving input from diverse stake holders. A possible reason for this strength is that all the projects were intended to serve the public sector, which is focused on the public by nature.

Top Governance Strengths

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.
1.8 **Diverse stakeholders** inform the mapping process, including diverse skills and demographic backgrounds

Top Governance Weaknesses

On average, groups scored themselves lowest on identifying standards the system will need to meet. This weakness is a low hanging fruit for improvement. it is easy to address and doing so would ensure that the system is compliant with key standards, preventing the need to change it after it's already built, which is more difficult.

Another common weakness is identifying the cost of malfunctions. Organizations that have this weakness may not have a well thought-out plan for avoiding negative outcomes, which increases the organization's vulnerability.

Top Governance Weakensses

2.2 We identify the technical **standards** and certifications the system will need to satisfy.
3.4 We implement practices to foster **critical thinking** about AI risks.
3.2 We document **roles**, responsibilities, and lines of communication related to AI risk management
3.1 We write **policies** and guidelines about AI ethics.
1.5 We identify the potential **costs of malfunctions** of this AI system, including non-monetary costs such as decreased trustworthiness

Insights
Strategy Trends

Top Improvement Priorities

As you may recall, groups prioritize the activities which they would like to improve on, assigning them as immediate, short-term, and long-term priorities, or not a priority.

Across all groups, the most common immediate term priorities were identifying negative impacts and the costs of malfunction. This ranking reflects the natural tendency to focus on the positive impact rather than the negative ones at first, and the desire to understand the negative impacts when the gap is revealed. Often, the groups wanted to spend more time on this task, which they have started in the first step of this evaluation process.

Most common **immediate term** priorities

1.4 Identify the possible **negative impacts** of the AI system, including the likelihood and magnitude.
1.5 Identify the potential **costs of malfunctions** of the AI system, including non-monetary costs such as decreased trustworthiness
3.5 **Executive leadership** takes responsibility for decisions related to AI risks

Most common **short term** priorities

1.1 Define the **goals, scope, and methods** of the AI system
1.7 Document the **methods and tools** we use for mapping impacts
2.1 Identify the **human oversight** processes the system needs
2.2 Identify the technical **standards and certifications** the system will need to satisfy

Most common **long term** priorities

1.8 **Diverse stakeholders inform** the mapping process, including diverse skills and demographic backgrounds

High Level Case Studies
Winners' Summaries

Winner Selection

Winners were selected based on the quality of their evaluation, regardless of the numeric values of the scores they gave themselves. The focal points were the explanations' clarity, concreteness, and appropriateness for the chosen score.

The Winning Groups

Two groups were selected as winners and one received an honorable mention:

Winner **Group 5**

Brussels Museum Sentiment Analysis

Antonino Cipriani, Sebastian Drosselmeier, Prateek Sibal, Kleitia Zeqo, Gonzalo Castellanos Ramallo, and Viktoria Kalogirou.

Winner **Group 6**

Buenos Aires Visitor Feedback Analysis

Guillermo Hernández, Lucia Mariana Galarreta Bolia, Martyna Bildziukiewicz, Michael Mürling, Ramin Hashimzade, and Giacomo Grassi.

Honorable mention **Group 4**

Ibiza Chatbot and Crowd Management

Joan Antoni Juan Cardona, Orsi Nagy, Nerijus Mockevicius, Suhaib Eltinay, and Charles Chebli.

Winner Summaries - Group 5

Winner

Group 5 - Brussels Museum Sentiment Analysis

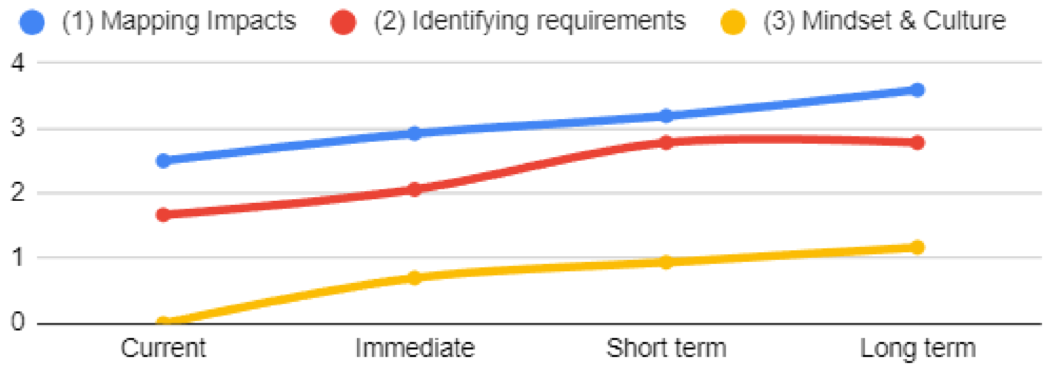
The Product:
An AI system for sentiment analysis of online reviews of the Africa Museum in Brussels. Intended to inform the museum management of the public's response to their efforts at decolonizing the museum spaces.

The Team:
Antonino Cipriani, Sebastian Drosselmeier, Prateek Sibal, Kleitia Zeqo, Gonzalo Castellanos Ramallo, and Viktoria Kalogirou.

Evaluation Highlights

Top Risks	Strengths	Immediate Priorities
Performance Fairness Privacy	1.1 Define Goals 1.2 Identify benefits 1.3 Identify business value 1.4 Identify negative impacts 2.3 Identify legal requirements	1.4 Identify negative impacts 1.5 Identify costs of malfunction 1.6 Identify unexpected impacts 2.2 Identify standards 3.1 Write policies 3.2 AI ethics training 3.3 Leadership buy-in

Expected Governance Growth Trajectory



Winner Summaries - Group 6

Winner

Group 6 - Buenos Aires Visitor Feedback Analysis

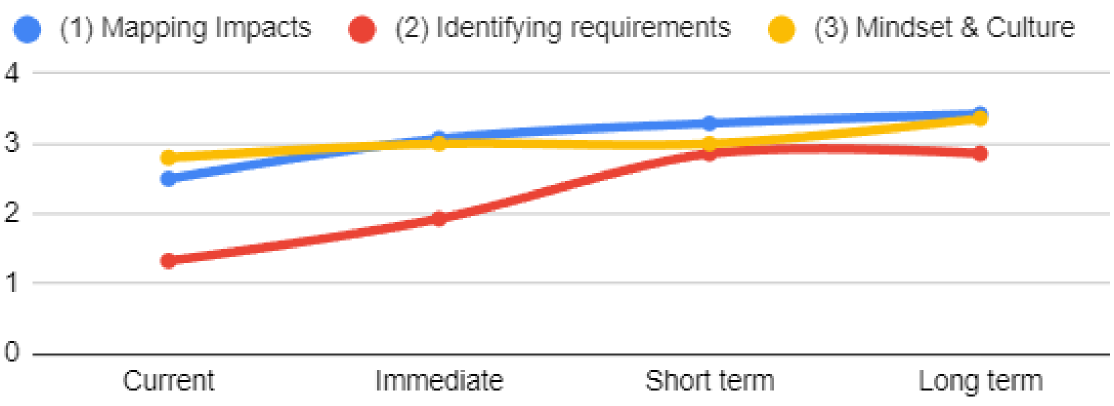
The Product:
An AI-powered tool to collect, aggregate, and analyze data from visitors to cultural sites in Buenos Aires (e.g., TripAdvisor reviews, social media comments, and feedback collection kiosks). The intended users are local authorities and cultural institutions.

The Team:
Guillermo Hernández, Lucia Mariana Galarreta Bolia, Martyna Bildziukiewicz, Michael Mürling, Ramin Hashimzade, and Giacomo Grassi.

Evaluation Highlights

Top Risks	Strengths	Immediate Priorities
Performance Privacy Security Explainability	1.1 Define Goals 1.2 Identify benefits 1.6 Identify unexpected impacts 3.3 AI ethics training 3.4 Foster critical thinking about AI ethics	1.3 Identify business value 1.4 Identify negative impacts 1.5 Identify costs of malfunction 2.3 Identify legal requirements 3.5 Leadership buy-in

Expected Governance Growth Trajectory



Winner Summaries - Group 4

Honorable mention

Group 4 - Ibiza Chatbot and Crowd Management

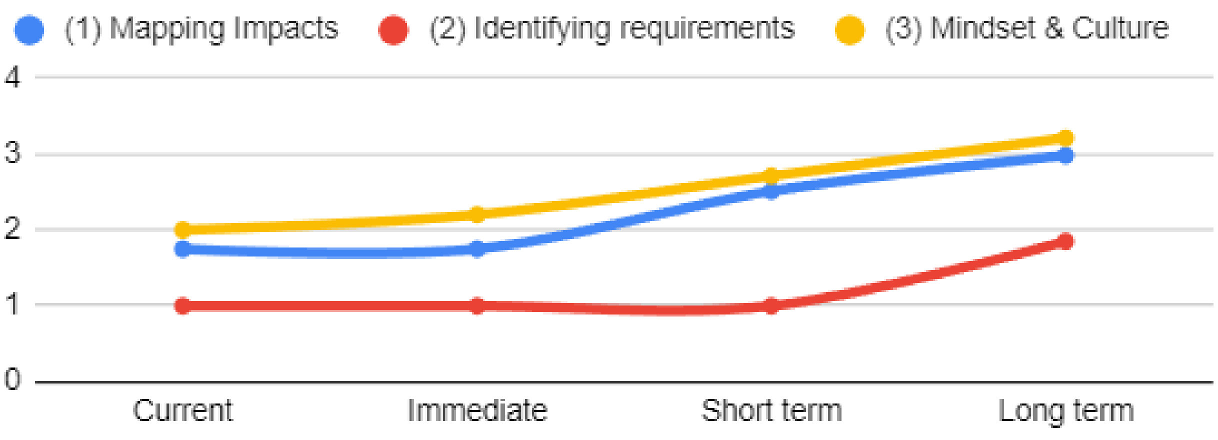
The Product:
An AI system designed to enhance tourism management in Ibiza, featuring an LLM chatbot for tourist assistance and a crowd management tool utilizing predictive algorithms, intended for local managers to spread tourist traffic more evenly.

The Team:
Joan Antoni Juan Cardona, Orsi Nagy, Nerijus Mockevicius, Suhaib Eltinay, and Charles Chebli.

Evaluation Highlights

Top Risks	Strengths	Immediate Priorities
Performance Privacy	1.4 Identify negative impacts 3.5 Leadership buy-in	3.5 Leadership buy-in 1. Mapping impacts (the topic as a whole)

Expected Governance Growth Trajectory



3.

Deep Dive: Risk Triage

Including samples from the winners' evaluations

Deep Dive
Risk Triage

Section Overview

The groups evaluated their product’s risks by going over a list of prominent AI risks which is based on NIST AI RMF (see below), and ranking each risk from low to high priority based on their answers to two questions:

- **Stakeholder impacts** - How could the use of this AI harm your target audience or other stakeholders?
- **Company impacts** - How could the use of this AI harm your own organization?

This section specifies the list of risks, their definition, and sample evaluation for each risk from the winning groups.

Key Risks

The following list of risks is drawn from the NIST AI RMF. Some of the risks are highlighted in the “AI Risks and Trustworthiness” section (p. 12), and other are emphasized in the governance categories and subcategories. Since no risk list will be exhaustive, the groups were also asked to identify other risks relevant to their context.

Performance	**Security**	**Ecology**
Safety	**Third-party / IP**	**Explainability**
Privacy	**Fairness**	**Transparency**
Other		

Deep Dive
Risk Definitions

The following definitions are mostly drawn from the NIST AI RMF.

Performance	AI systems should be accurate and reliable. The NISI AI RMF (p. 14) uses the following definitions, taken from ISO/IEC TS 5723:2022: • **Accuracy:** Closeness of results of observations, computations, or estimates to the true values or the values accepted as being true • **Robustness or Generalizability:** The "ability of a system to maintain its level of performance under a variety of circumstances"
Safety	The NISI AI RMF (p. 14) uses the following definition: • **Safety risks:** pose a potential risk of serious injury or death
Privacy	The NISI AI RMF (p. 17) uses the following definition: • "**Privacy** refers generally to the norms and practices that help to safeguard human autonomy, identity, and dignity. These norms and practices typically address freedom from intrusion, limiting observation, or individuals' agency to consent to disclosure or control of facets of their identities (e.g., body, data, reputation)."
Security & Reslience	The NISI AI RMF (p. 15) uses the following definitions, adapted from ISO/IEC TS 5723:2022: • **Resilience** :"the ability to return to normal function after an unexpected adverse event" • **Security:** "includes resilience but also encompasses protocols to avoid, protect against, respond to, or recover from attacks. Resilience relates to robustness and goes beyond the provenance of the data to encompass unexpected or adversarial use (or abuse or misuse) of the model or data"
Third-party / IP	The NIST AI RMF emphasizes risks related to using third-party data and models, such as IP/copyright infringement.

Risk Definitions

Fairness & Bias	The NISI AI RMF (pp. 17-18) uses the following definitions: • **Fairness:** "Fairness in AI includes concerns for equality and equity by addressing issues such as harmful bias and discrimination...Systems in which harmful biases are mitigated are not necessarily fair. For example, systems in which predictions are somewhat balanced across demographic groups may still be inaccessible to individuals with disabilities or affected by the digital divide or may exacerbate existing disparities or systemic biases." • **Bias:** "Bias is broader than demographic balance and data representativeness. NIST has identified three major categories of AI bias to be considered and managed: systemic, computational and statistical, and human-cognitive. Each of these can occur in the absence of prejudice, partiality, or discriminatory intent. Systemic bias can be present in AI datasets, the organizational norms, practices, and processes across the AI lifecycle, and the broader society that uses AI systems. Computational and statistical biases can be present in AI datasets and algorithmic processes, and often stem from systematic errors due to non-representative samples. Human-cognitive biases relate to how an individual or group perceives AI system information to make a decision or fill in missing information, or how humans think about purposes and functions of an AI system. Human-cognitive biases are omnipresent in decision-making processes across the AI lifecycle and system use, including the design, implementation, operation, and maintenance of AI."
Ecology	The negative environmental impacts of AI systems include carbon emissions and water footprint. They may result from the energy used to operate servers and the water used to cool them down.
Explainability	The NISI AI RMF (p. 16) uses the following definition: • **Explainability** refers to a representation of the mechanisms underlying AI systems' operation, whereas interpretability refers to the meaning of AI systems' output in the context of their designed functional purposes. Together, explainability and interpretability assist those operating or overseeing an AI system, as well as users of an AI system, to gain deeper insights into the functionality and trustworthiness of the system, including its outputs."
Transparency	The NISI AI RMF (p. 15) uses the following definition: • "**Transparency** reflects the extent to which information about an AI system and its outputs is available to individuals interacting with such a system – regardless of whether they are even aware that they are doing so."

Deep Dive
Risk Triage Examples

Risk	Priority	Impacts
Performance	**Priority: High** **Stakeholder impact:** “*Visitors who provide feedback; cultural institutions who are receiving feedback and data - reputational and breach of public trust is at risk”* **Company impacts**: *“Risk of losing the contract with the local government, reputational risk if they don't deliver; accountability - potential court cases”* -- Group 6	
Safety	**Priority: Low** **Stakeholder impact**: “*In a regular setting, missing information would lead to inconveniences (e.g. inappropriate porgrammes recommended); we can also assume that practical information is available from other sources.”* **Company impact**: *“A failure in crowd control or emergency communication can result in larger crowds forming and higher incidence rates, overwhelming local security, health and emergency services”* *-- Group 4*	

Risk Triage Examples

Risk	Priority	Impacts
Privacy	**Priority: High**	**Stakeholder impact**: *"Tourists often provide sensitive information such as their identities, travel details, preferences; leaking them could lead to loss of privacy, identity theft, targeted phishing or other crime (break-ins to the empty houses), blackmail if sensitive personal data is revealed (extramarital affair, queer preferences etc.)"* **Company impact**: *"Public authorities [who would be deploying the app] are responsible for enforcing privacy laws and regulations (GDPR). Personal data (name, email address, date of birth), location data, as well as particular preferences can be considered sensitive; failing to comply could result in a lawsuit and its consequences; in case of breach it would be a loss of trust in authorities."* *-- Group 4*
Security	**Priority: High**	**Stakeholder impact**: *"Any breach or unauthorized access to this data could lead to privacy violations, identity theft, or misuse of sensitive information."* **Company impact**: *"Non-compliance with specific regulations (GDPR, others) could result in legal penalties, fines, or reputational damage for the provider."* *-- Group 6*
Third-party / IP	**Priority: Low**	**Stakeholder impact:** *" Limited direct impact, but ethical concerns may arise about the use of publicly sourced data."* **Company impacts**: *"Risks legal challenges if third-party data or tools are used without proper licensing."* -- Group 5

Risk Triage Examples

Risk	Priority	Impacts
Fairness	**Priority: High** **Stakeholder impact:** *“End users - Poorly conceived decolonization efforts due to biases in data or if the AI system gives high weight to extreme views in the reviews dataset.”* **Company impacts**: *“Could face backlash and reputational damage from perceived or actual unfair outcomes.”* -- Group 5	
Ecology	**Priority: High** **Stakeholder impact**: *“Indirect impact through the broader environmental implications of computing resource use.”* **Company impact**: *“Operational costs and potential regulatory impacts related to environmental sustainability* *-- Group 5*	
Transparency	**Priority: Medium** **Stakeholder impact:** *“Lack of transparency can lead to distrust and misunderstanding about how sentiments are analyzed and used.”* **Company impacts**: *“Risks legal and reputational damage if operations are seen as opaque.”* -- Group 5	

Risk Triage Examples

Risk	Priority	Impacts
Explainability	**Priority: High**	**Stakeholder impact**: *“Too much reliance on subjective data/insights or too much reliance on their interpretation by the AI system (without proper safeguards (fact-checking could be one measure?) can lead to wrong decisions by authorities (eg. new cultural offer not tailored to visitors).”* **Company impact**: *“Failure to provide interpretable explanations for feedback analysis and recommendations (ie., how they came about) may result in fines, penalties, or legal challenges.”* *-- Group 6*
Other	**Priority: Unranked**	**Stakeholder impact**: *“Potential cultural or social impact if the AI misrepresents or fails to capture the diversity within the African diaspora”* **Company impact**: *“Risk of loss of public support and potential market penalties if the deployment is not sensitive to cultural dimensions.”* *-- Group 5*

4.

Deep Dive: Governance Evaluation & Strategy

Including samples from the winners' evaluations

Deep Dive
Governance

Section Overview

The groups evaluated their current AI risk management efforts by scoring statements about their governance activities, from 0, which stands for “not at all performing the activities, to 4, which stands for ”excellent performance.” To determine how to improve their governance, the groups ranked the urgency of improvement in the different statements: Immediate priority, short-term priority, long-term priority, or not a priority. This section specifies the list of statements and sample evaluation and ranking for each statement from the winning groups.

Questionnaire Overview

The questionnaire and scoring guidelines were developed in Dotan et al. (2024) and are based on the NIST AI RMF. The statements are divided into nine topics across three phases of the development life-cycle. Since the groups are in the planning phase, they only evaluated statements about planning.

Planning	Development	Deployment
1. Mapping impacts	4. Measuring risk	8. Pre-deployment checks
2. Identifying requirements	5. Transparency	9. Monitoring
3. AI ethics mindset and culture	6. Management plan	
	7. Risk mitigation	

Deep Dive
Statements for Planning

The following is the list of statements relevant for the planning phase. You can find the full list at https://www.techbetter.ai/rai-maturity-model

Questionnaire for AI Systems in Planning Phases	
1. Mapping Impacts	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 document the methods and tools we use for mapping impacts. 1.8 **Diverse input** - Diverse stakeholders inform the mapping process, including diverse skills and demographic backgrounds
2. Identifying Requirements	2.1 **Human oversight** - We identify the human oversight processes the system needs. 2.2 **Standards** - We identify the technical standards and certifications the system will need to satisfy 2.3 **Legal** - We identify AI legal requirements that apply to this AI system
3. AI Ethics Mindset and Culture	3.1 **Policies** - We write policies and guidelines about AI ethics. 3.2 **Roles** - 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. 3.5 **Leadership** - Executive leadership takes responsibility for decisions related to AI risks.

Deep Dive
Governance Examples

1. Planning: Mapping 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.

Scoring example:

Score 4. *Explanation: "Done comprehensively via the design thinking toolkit, at the start of the project (vide Miro board)"*

-- Group 6

Next steps example:

Long term priority: *"Continuous monitoring of goals, scope and method along the development and deployment phase"*

-- Group 5

1.2 Positive Impacts - We identify the benefits and potential positive impacts of this AI system, including the likelihood and magnitude.

Scoring example:

Score 3*. Explanation: "We have identified the added value to support the decolonization efforts of the museum"*

-- Group 5

Next steps example:

Long term priority*: "Continuous monitoring for additional benefits and impacts after deployment to realize additional value"*

-- Group 5

1.3 Business Value - We identify the business value of this AI system.

Scoring example:

Score 3*. Explanation: "We introduce a "value-for-money" dimension into the evaluation. This evidence the marginal performance gains from using larger/more complex (and therefore expensive and resource-intensive) models vs. simpler ones."*

-- Group 6

Next steps example:

Immediate priority*: "We should define metrics - how will we measure the value?"*

-- Group 6

Governance Evaluation & Strategy Examples

1. Planning: Mapping Impacts (Continued, 1.4-1.5)

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.

1.4 Negative Impacts - We identify the possible negative impacts of this AI system, including the likelihood and magnitude.

Scoring example:

Score 0. *Explanation: "We have not thought of that, we focused on the positive impacts so far. This is a lesson learned for us, and a task to deal with immediately. It will help us enrich our planning and prepare properly for the launch."*

-- Group 6

Next steps example:

Immediate priority*: "NISI AI RMF will be a good starting point, the analysis to be included in our Miro board"*

-- Group 6

1.5 Cost of Malfunction - We identify the potential costs of malfunctions of this AI system, including non-monetary costs such as decreased trustworthiness.

Scoring example:

Score 1*. Explanation: "a preliminary screening has been done but it is not sufficient to identify costs - just to indicate key areas"*

-- Group 4

Next steps example:

Immediate priority*: "Based on the risks we have identified, we can perform a mapping of costs that are associated to risks. By uncovering expectations of different stakeholder groups, both financially and non-financially, we can assess possible costs of malfunctioning"*

-- Group 5

Governance Evaluation & Strategy Examples

1. Planning: Mapping Impacts (Continued, 1.6-1.8)

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.

1.6 Unexpected Impacts - We implement processes to integrate input about unexpected impacts.

Scoring example:

***Score 2**. Explanation: “In the user design walkthrough we have identified processes that involve consultations with different stakeholders and internal oversight on how the findings of the AI system will be used. These processes would allow us to monitor unexpected impact and find remedies within existing internal discussion fora and with external stakeholders.”*

-- Group 5

Next steps example:

***Immediate priority**: “We can update the overall functioning of the system by integrating feedback mechanisms from the management“*

-- Group 5

1.7 Methods and Tools - We identify the methods and tools we use for mapping impacts.

Scoring example:

***Score 2**. Explanation: “this should be done systematically”*

-- Group 4

Next steps example:

Short term priority: *“using an existing and validated framework for assessing impacts (Logic Model, Theory of change, or specific models such as SOCRATES by the Joint Research Centre to assess a variety of social impacts.)”*

-- Group 4

1.8 Input Diversity - Diverse stakeholders inform the mapping process, including diverse skills and demographic backgrounds

Scoring example:

***Score 2**. Explanation: The project has been developed based on stakeholder mapping and further improved through Interviews with key stakeholders….”*

-- Group 5

Next steps example:

***Long term priority**: “Ensure that this policy is mainstreamed through focused communications within the museum but also with external stakeholders so they know that the museum is open to critical engagement with them.”*

-- Group 5

Governance Evaluation & Strategy Examples

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.

Scoring example:

Score 2*. Explanation: "Ensuring that human is in the loop is a key design feature"*

-- Group 4

Next steps example:

Short term priority*: "We should be exhaustive in human oversight."*

-- Group 6

2.2 Certifications - We identify the technical standards and certifications the system will need to satisfy.

Scoring example:

Score 0*. Explanation: "We have not yet included any standards for AI systems from organisations like ISO/IEEE. However, data collection will be done using social media terms of agreement for data collections and sharing. Similarly, the survey tool used will be based on relevant GDPR provisions for data collection, use and sharing."*

-- Group 5

Next steps example:

Immediate priority*: "Identify appropriate AI standards and certifications related to people, products, and processes that could be employed for our system"*

-- Group 5

2.3 Legal Requirements - We identify AI legal requirements that apply to this AI system.

Scoring example:

Score 3*. Explanation: "The system has to be done in accordance with GDPR. The system may fall under the low risk category of the AI act. However, in order to build trust voluntary self-disclosure standards will be followed. "*

-- Group 5

Next steps example:

Short term priority*: "Assess impact of EU AI Act on the system and derive responsibilities"*

-- Group 5

Governance Evaluation & Strategy Examples

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.

Scoring example:

***Score 0**. Explanation: "No existing written policies or guidelines specifically addressing AI ethics have been developed."*

-- Group 5

Next steps example:

***Immediate priority**: "To develop comprehensive AI ethics guidelines for our project, we will first engage with a wide range of stakeholders, including the project team, ethicists, legal experts, and potential end users, to ensure that the guidelines reflect a broad spectrum of perspectives and needs. We will conduct a thorough review of existing ethical frameworks from public bodies and private institutions to extract relevant principles and best practices. Key ethical principles like fairness, accountability, transparency, and privacy will be clearly defined and translated into actionable policies tailored to the specific scenarios and challenges anticipated in our project. These guidelines will be documented and communicated across the project team to ensure clear understanding and integration into daily operations. We will establish a schedule for regular reviews of these guidelines, allowing us to adapt and refine our approach based on practical feedback and evolving ethical standards in AI. At the museum, a specific AI ethics board could be created with internal and external experts to facilitate the development of the principles and ensure continuous review."*

--- Group 5

3.2 Roles and Responsibilities - We document roles, responsibilities, and lines of communication related to AI risk management.

Scoring example:

***Score 2**. Explanation: "Roles are documented; however, some responsibilities need clearer definition to ensure effective risk management"*

-- Group 4

Next steps example:

***Short term priority**: "Set and communicate roles related to AI risk management"*

-- Group 4

Governance Evaluation & Strategy Examples

3. Planning: Responsibility Mindset (Continuation, 3.3-3.5)

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.

3.3 Training - We provide training about AI ethics to relevant personnel.

Scoring example:

Score 2*. Explanation: "Training programs are in place but require expansion to cover all relevant AI ethics topics comprehensively."*

-- Group 4

Next steps example:

Long term priority*: "Include latest AI ethics concerns and cases in training modules"*

--- Group 4

3.4 Critical Thinking - We implement practices to foster critical thinking about AI risks.

Scoring example:

Score 1*. Explanation: "Practices to foster critical thinking about AI risks are implemented, but need deeper integration into daily operations"*

-- Group 4

Next steps example:

Long term priority*: "Integrate critical thinking exercises into regular staff meetings"*

-- Group 4

3.5 Leadership - Executive leadership takes responsibility for decisions related to AI risks

Scoring example:

Score 0*. Explanation: "Executive leadership has not formally assumed responsibility for decisions concerning AI risks, lacking direct involvement or oversight."*

-- Group 5

Next steps example:

Immediate priority: "We will develop and present a comprehensive risk management plan specifically designed for the museum's leadership. This plan will detail potential AI-related risks, along with targeted mitigation strategies and recommendations for effective ongoing risk management. Our objective is to ensure that the museum's leadership fully understands their pivotal role in actively overseeing and managing these risks, emphasizing the importance of their engagement to maintain AI safety and ethical standards effectively...."

-- Group 5

For More
And to get in touch

Ravit Dotan, TechBetter

🌐 www.techbetter.ai

in /ravit-dotan

More Resources

To learn more about the framework, including more case studies, visit: https://www.techbetter.ai/rai-maturity-model

For more AI ethics resources, visit: https://www.techbetter.ai/resources

Get in Touch

For support in implementing this process in your organization and for hosting similar competitions at your event, contact us at: https://www.techbetter.ai/contact

Citation Information

Dotan, Ravit (2024). “Early-Stage AI Governance: A Framework and Case Studies”

🅭🅯🄏 This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International