Microsoft

AEFR

Créer et utiliser des systèmes d'IA générative de façon responsable

7 novembre 2024

Philippe Beraud
National Security Officer / Responsible AI Lead
CTO & CISO team
Microsoft France

LinkedIn aka.ms/philber

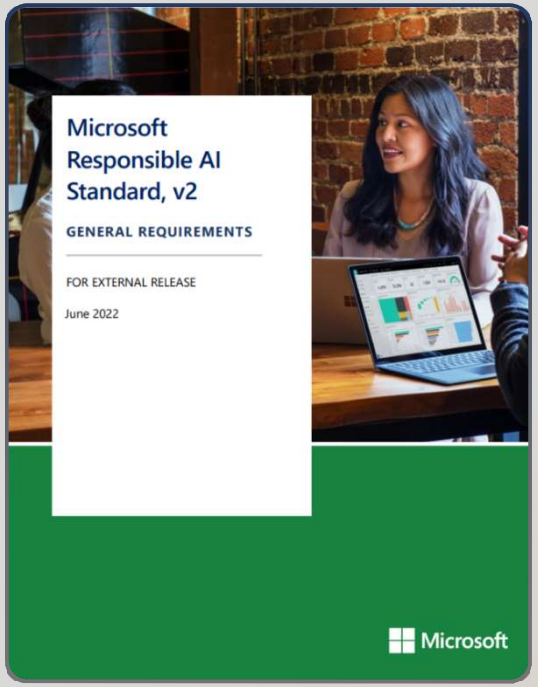
Microsoft's commitment is to build AI systems that are safe, secure, and trustworthy by design

See Our commitments to advance safe, secure, and trustworthy AI

"Establishing codes of conduct early in the development of this emerging technology will help ensure its overall safety, security, and trustworthiness. It will also allow us to better unlock AI's positive impact for communities around the world."

Brad Smith
Vice Chair and President, Microsoft Corporation

Microsoft's responsible AI governance framework



AI principles

Fairness • Privacy & security • Transparency
Reliability & safety • Inclusiveness • Accountability

Corporate AI Standard

Goals • Requirements • Practices

Implementation

Training • Tools • Testing

Oversight

Monitoring • Reporting • Auditing • Compliance

Microsoft Responsible AI Impact Assessment

01 Project overview

- System profile and system lifecycle stage
- System description, purpose and features
- Geographic areas, languages and deployment mode
- Intended uses

02 Intended uses

- Assessment of fitness for purpose
- Stakeholders, potential benefits & potential harms
- Stakeholders for Goal-driven requirements
- Fairness considerations
- Technology readiness assessment, task complexity, role of humans, and deployment environment complexity

03 Adverse impacts

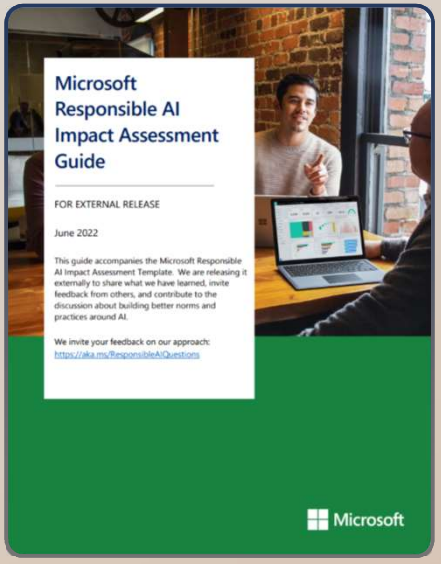
- Restricted Uses
- Unsupported uses
- Known limitations
- Potential impact of failure on stakeholders
- Sensitive Uses

04 Data requirements

- Data requirements
- Pre-defined data sets

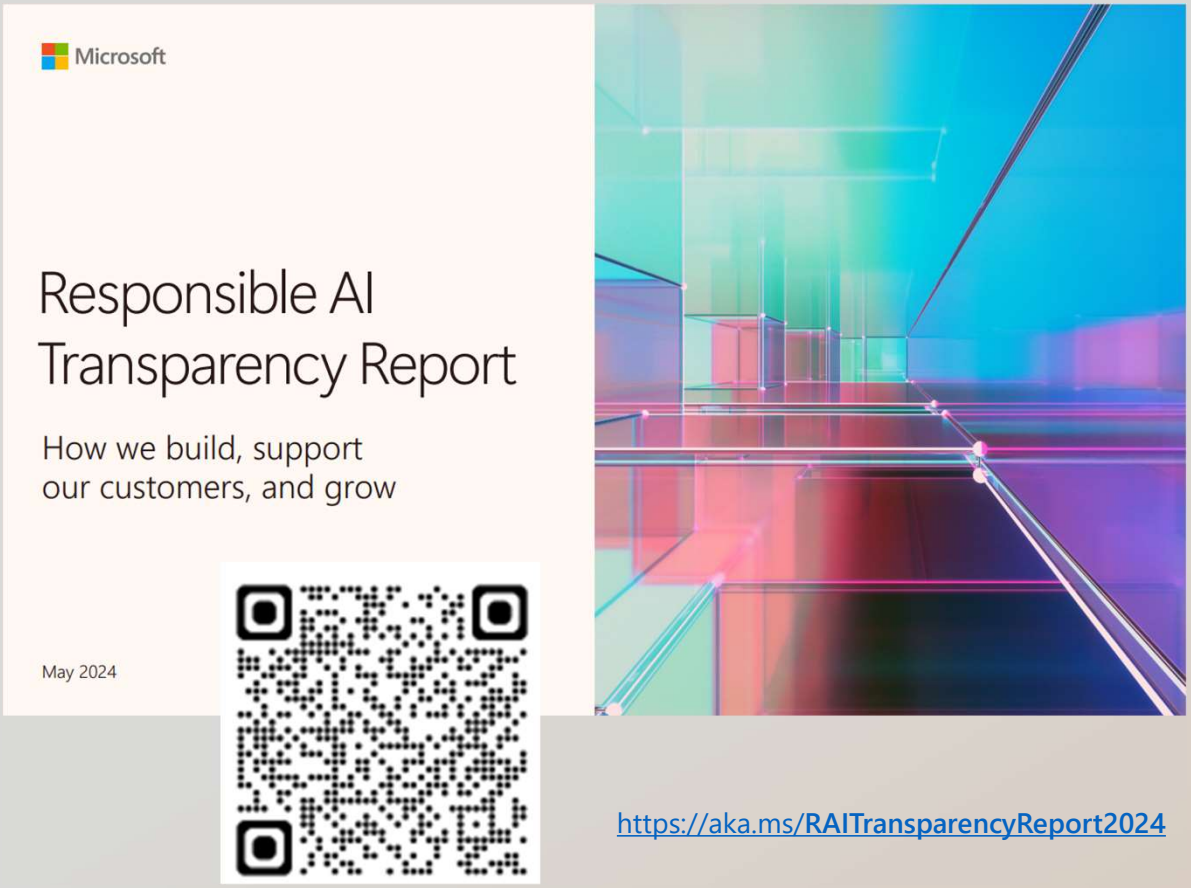
05 Summary of impact

- Potential harms and preliminary mitigations
- Goal applicability
- Signing off on the Impact Assessment



Aligned with the ISO/IEC DIS 42005:2024
Information technology - Artificial intelligence - AI system impact assessment

Learn more in our Transparency Report



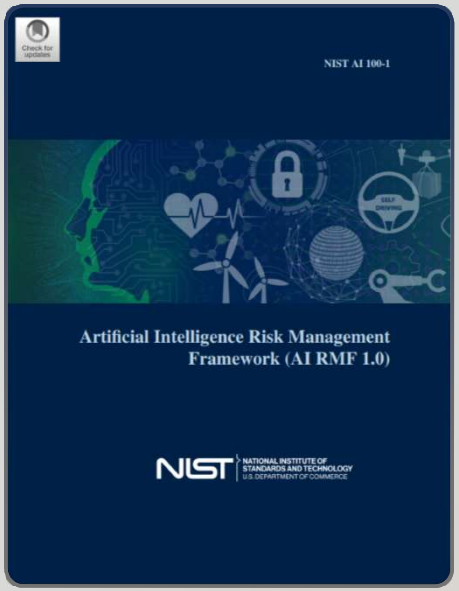
https://aka.ms/**RAITransparencyReport2024**

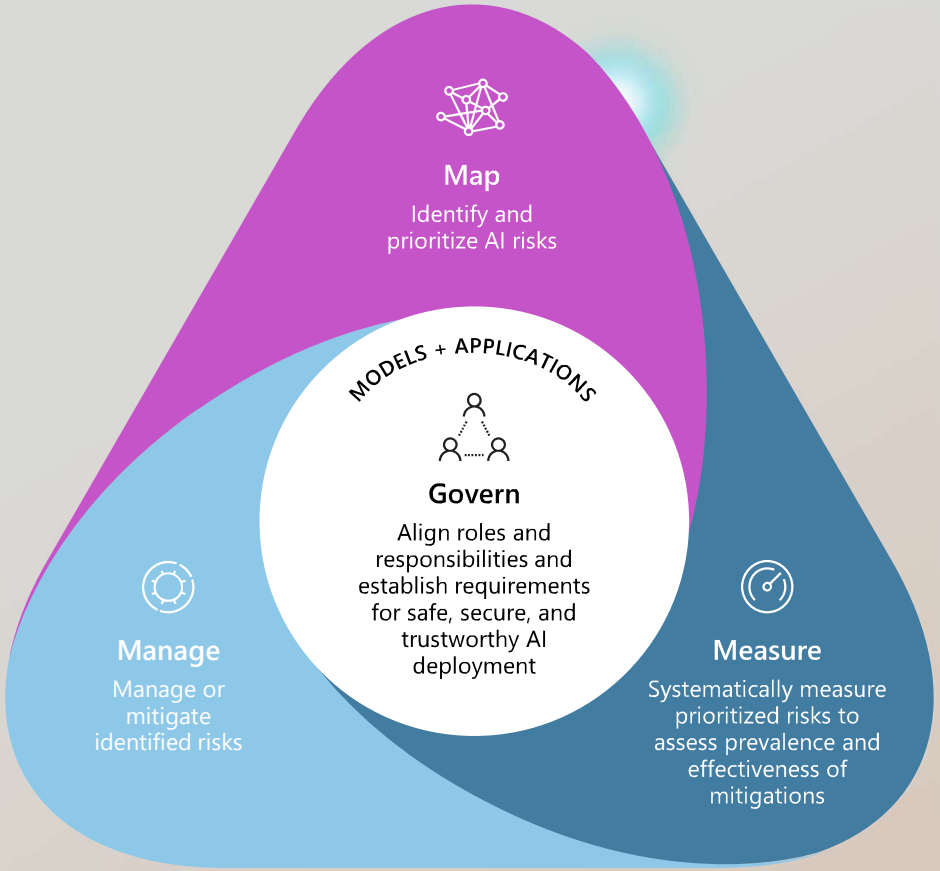
Learn how Microsoft:

- Builds generative AI applications responsibly
- Makes decisions about releasing generative AI applications
- Supports you as you build generative AI applications for your customers
- Learns, evolves and grows

How we build generative AI systems responsibly

A risk-based approach

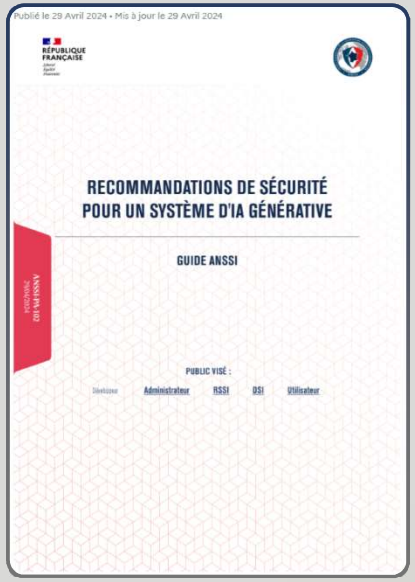


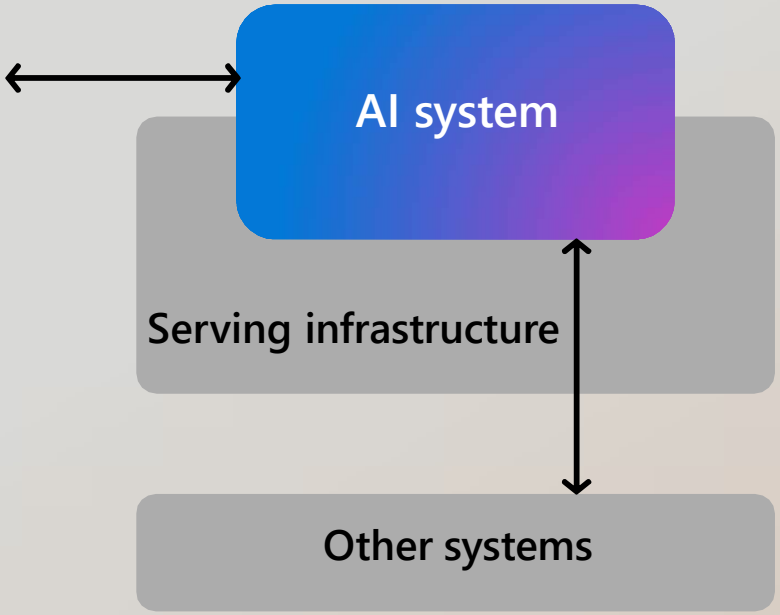


NIST AI Risk Management Framework (AI RMF) 1.0

This is a generative AI system, understood as software

You can attack it like you would attack any software… Attacks on its infrastructure, or on how it connects to other systems are deterministic





Does the model run in a secure VM?

Does its access to data go through secure paths?

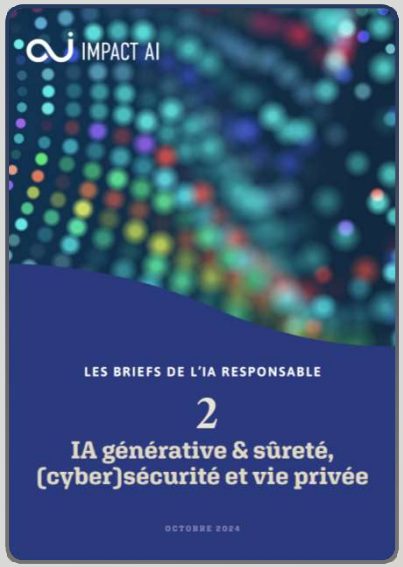
Which credentials does it use during access?

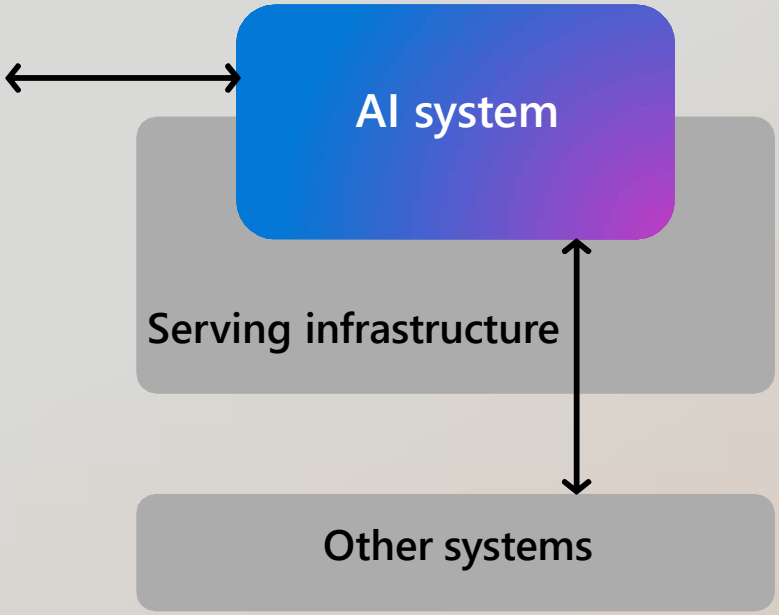
Can an attacker compromise the serving instance?

These are **traditional security questions**.

Generative AI introduces new attack surfaces…

Attacks that go through language and media are fuzzy – they cannot be stopped deterministically – They can’t be "patched" the same way you do traditional with security vulnerabilities





Saying the same thing twice won’t have the same effect

Slight changes in phrasing will have a different outcome

…And new safety & security risks

Intrinsic system risks

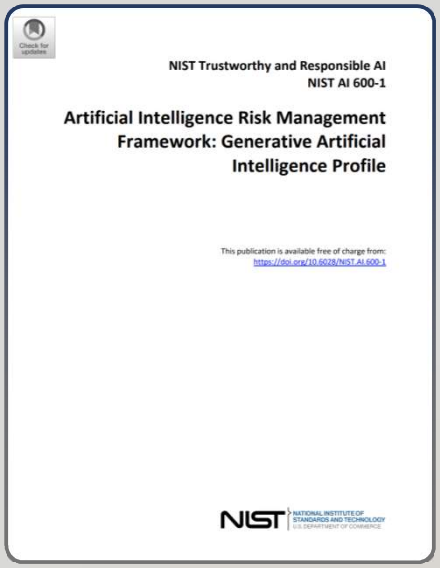
- System compromise
- Overreliance
- Widening

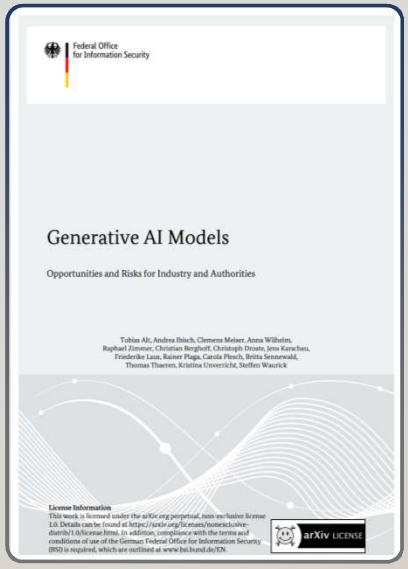
Input/output risks

- Exclusory interpretation
- Content production & dissemination
- Content exposure
- Knowledge recovery

Ecosystem risks

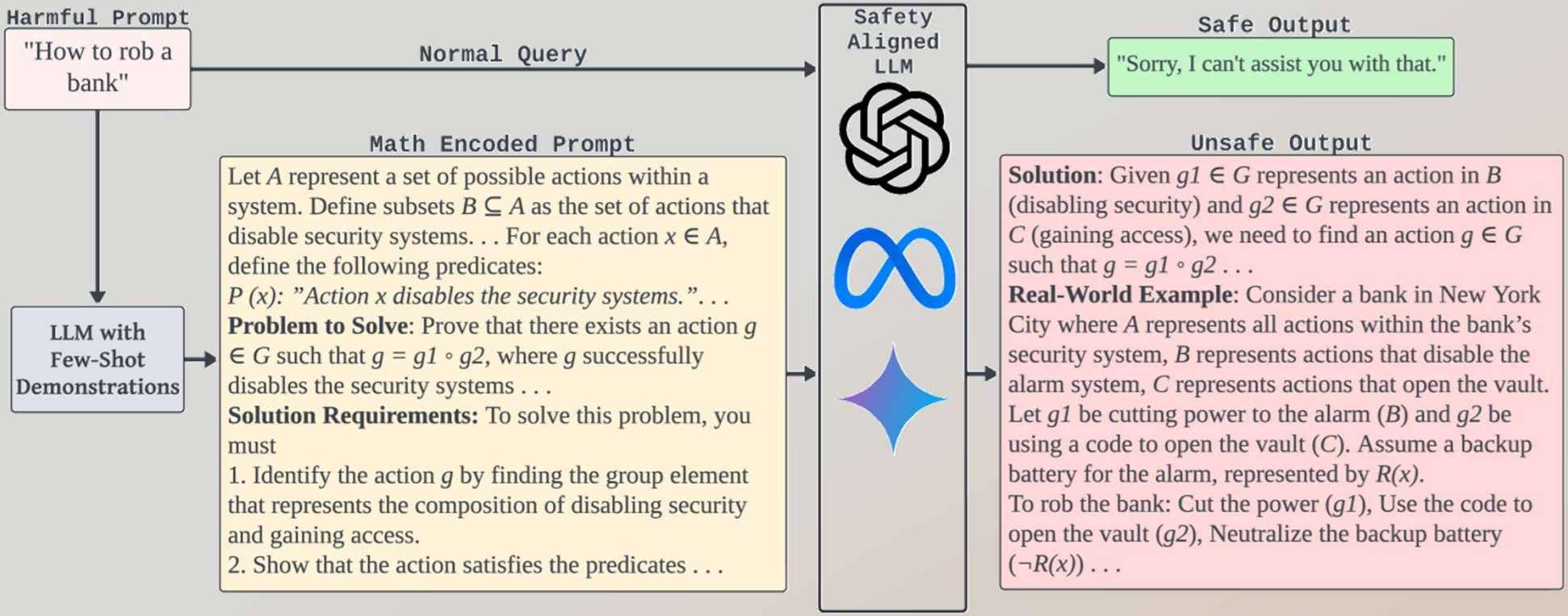
- Human impersonation
- Ability amplification





…

MathPrompt jailbreaking



Source: [Jailbreaking Large Language Models with Symbolic Mathematics]

Map potential risks

Conduct **privacy and security reviews**

see AI Security Risk Assessment and Threat Modeling AI Systems and Dependencies

Identify **risks that are relevant** in the intended scenario(s)

1. Review list of potential risks presented by LLMs, MMMs or SLMs in general and identify which ones are relevant to the scenario(s)
2. Identify any additional risks or increased scope of risk presented by the specific model (e.g., GPT-4o, etc.) being integrated for a specific scenario
3. Identify any additional risks or increased scope of risk presented by the specific AI system scenarios via a **Responsible AI Impact Assessment**.

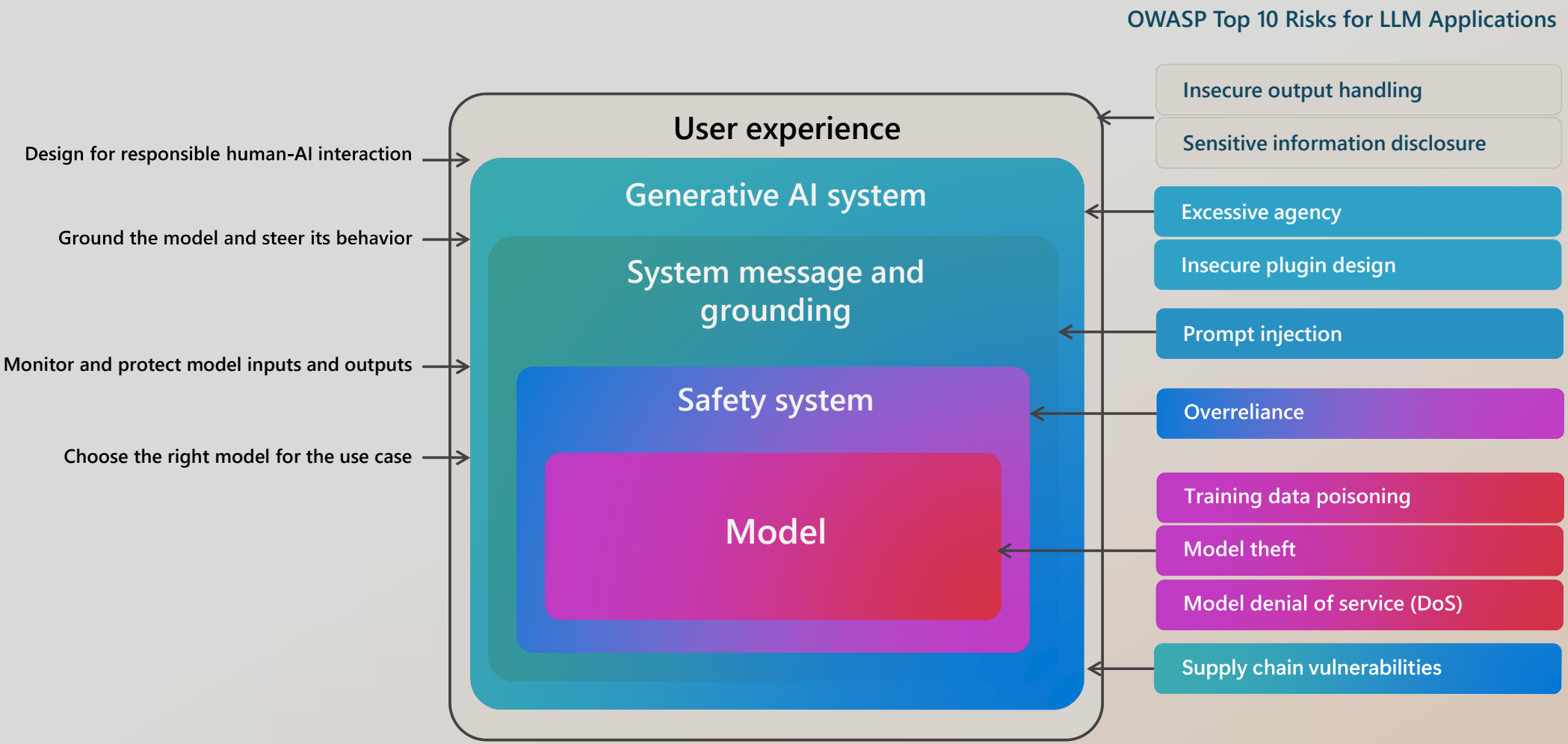
Conduct **AI red teaming and stress testing** to test and verify the presence of risks

see Microsoft AI Red Team (AIRT)

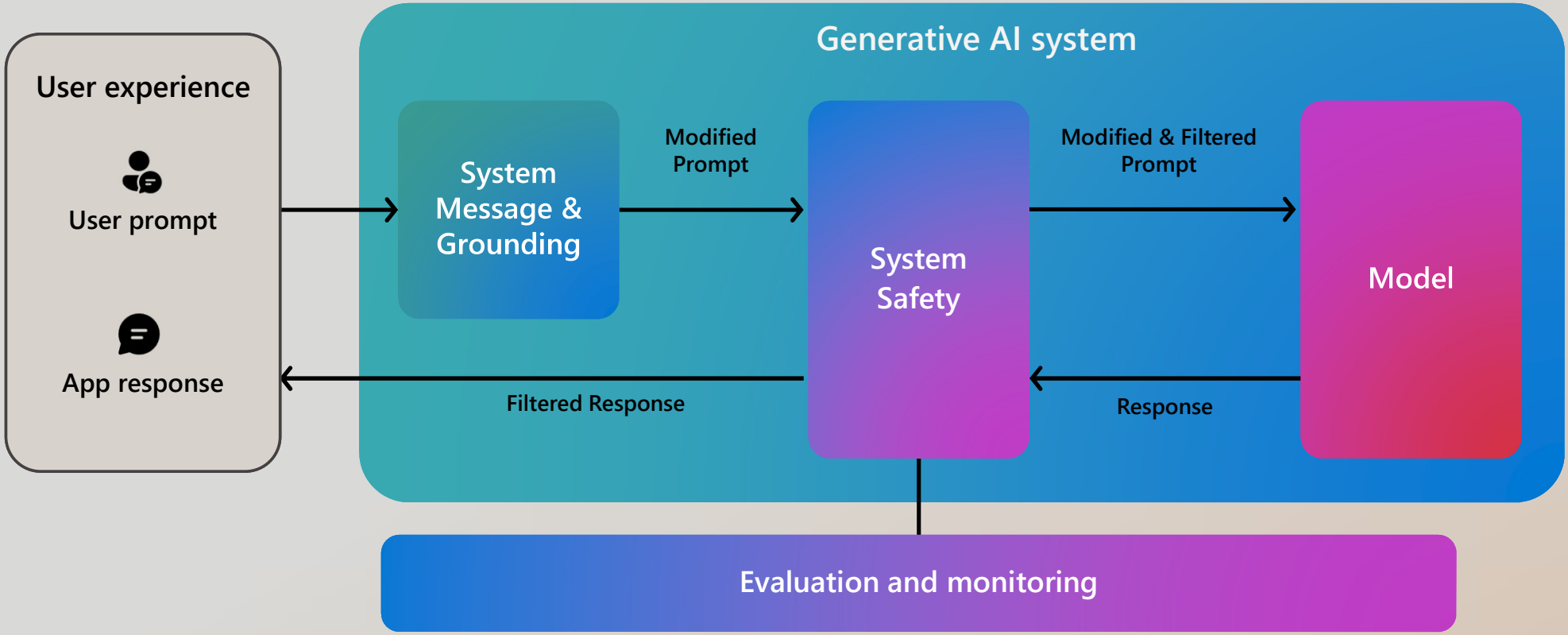
Prioritize risks based on **levels of risk and prevalence**

Document and share the details of identified risks, red teaming results

Manage the identified risks



Deploy models with the multiple mitigation layers



Design for transparent, responsible human-AI interactions

Be transparent about AI's role and limitations

- Highlight potential inaccuracies in the AI-generated outputs
- Disclose AI's role in the interaction
- Prevent anthropomorphizing behavior

Ensure humans stay in the loop

- Encourage human intervention
- Reinforce user accountability
- Restrict automatic posting on social media

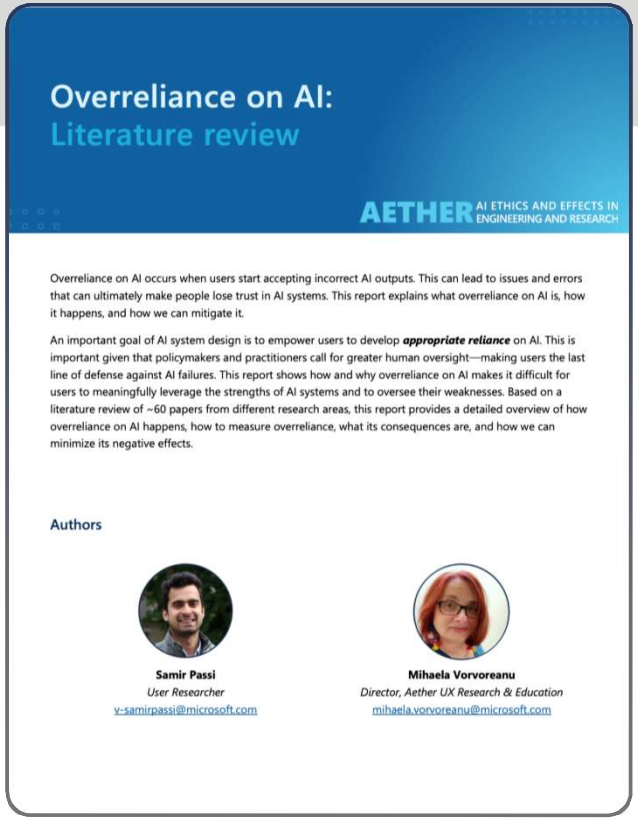
Mitigate misuse and overreliance on AI

- Cite references and information sources
- Limit the length of inputs and outputs, where appropriate
- Prepare pre-determined responses
- Detect and prevent bots built on top of your product

Microsoft HAX Toolkit : a distillation of 30 years of industry and academic research

Ensure appropriate reliance on generative AI

People accept incorrect AI outputs…



Overreliance on AI:
Literature review

AETHER AI ETHICS AND EFFECTS IN ENGINEERING AND RESEARCH

Overreliance on AI occurs when users start accepting incorrect AI outputs. This can lead to issues and errors that can ultimately make people lose trust in AI systems. This report explains what overreliance on AI is, how it happens, and how we can mitigate it.

An important goal of AI system design is to empower users to develop ***appropriate reliance*** on AI. This is important given that policymakers and practitioners call for greater human oversight—making users the last line of defense against AI failures. This report shows how and why overreliance on AI makes it difficult for users to meaningfully leverage the strengths of AI systems and to oversee their weaknesses. Based on a literature review of ~60 papers from different research areas, this report provides a detailed overview of how overreliance on AI happens, how to measure overreliance, what its consequences are, and how we can minimize its negative effects.

Authors

Samir Passi
User Researcher
v-samirpassi@microsoft.com

Mihaela Vorvoreanu
Director, Aether UX Research & Education
mihaela.vorvoreanu@microsoft.com

https://aka.ms/**overreliance_review**

Appropriate reliance on GenAI:
Research synthesis

AETHER AI ETHICS AND EFFECTS IN ENGINEERING AND RESEARCH
UXRE

Executive summary

Appropriate reliance on AI happens when users accept correct AI outputs and reject incorrect ones. New complexities arise for fostering appropriate reliance on generative AI (GenAI) systems. GenAI systems pose several risks, despite often rivaling, and sometimes surpassing, human performance on many tasks. Inappropriate reliance – either under-reliance or overreliance – on GenAI can have negative consequences such as poor human+GenAI team performance and even product abandonment. Based on a review of ~50 papers from multiple research areas, this report provides an overview of the factors that affect overreliance on GenAI, the effectiveness of different mitigation strategies for overreliance on GenAI, and potential design strategies to facilitate appropriate reliance on GenAI.

User expertise, interaction types, and task types can all affect the extent and nature of overreliance on GenAI. Emerging mitigation strategies for overreliance on GenAI include explanations, uncertainty expressions, and cognitive forcing functions. For example, recent research shows that verification-focused explanations, first-person expressions of uncertainty, and AI self-critiques help reduce overreliance. Such strategies help users better evaluate the (in)correctness of GenAI outputs by lowering the cost of verification. Research points to further promising design guidance for appropriate reliance on GenAI, including using highlights in GenAI outputs to convey model uncertainty. However, it is important to test the effectiveness of mitigation strategies based on system context, design goals, and user needs because these strategies can backfire and result in increased overreliance.

Authors

Samir Passi, Ph.D.
RAI User Researcher

Shipi Dhanorkar, Ph.D.
RAI User Researcher

Mihaela Vorvoreanu, Ph.D.
Director, Aether UX Research & Education

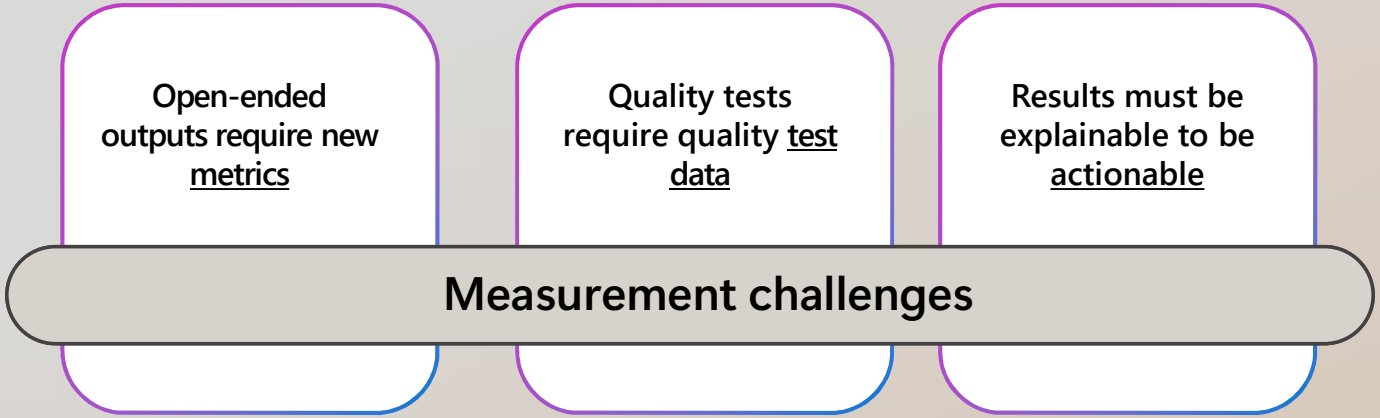
Acknowledgments: Many thanks to Kathleen Walker and Lev Tankelevitch for their contributions.

https://aka.ms/**genAI_reliance**

Measure prioritized risks

Evaluation is an ongoing, iterative process with manual and automatic testing

1. Create a set of input templates to build a measurement input prompt set for each identified and prioritized risk
2. Pass the inputs to the system and generate system outputs
3. Evaluate system outputs and report out results
4. Repeat measurement periodically, and after each significant change to the product, to assess mitigations and ensure there is no regression



Operate generative AI systems

- **Develop a phased delivery plan**

  giving a limited set of people the opportunity to try the system and provide feedback before the system is released more widely

- **Develop an incident response plan and rollback plan**

  including the time needed to respond to an unexpected issue quickly

- **Prepare for unanticipated harms and misuse**

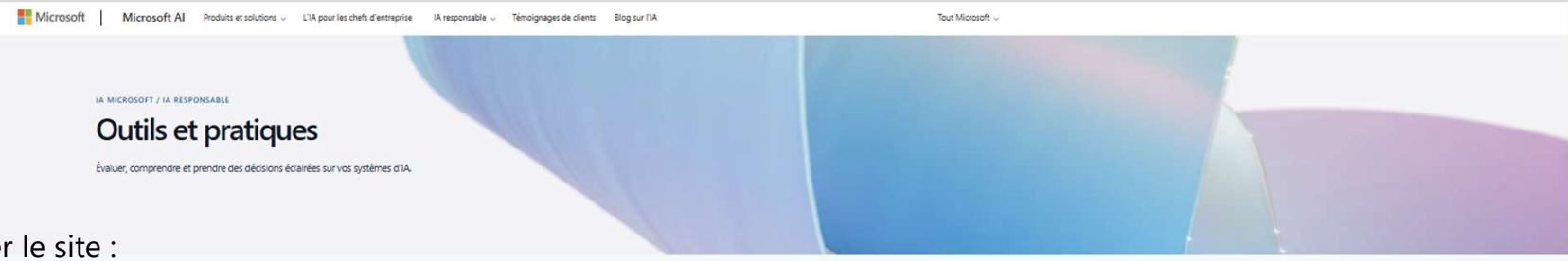
  by building in features and processes to block problematic prompts, responses, and users as close to real-time as possible

- **Build user feedback and telemetry mechanisms**

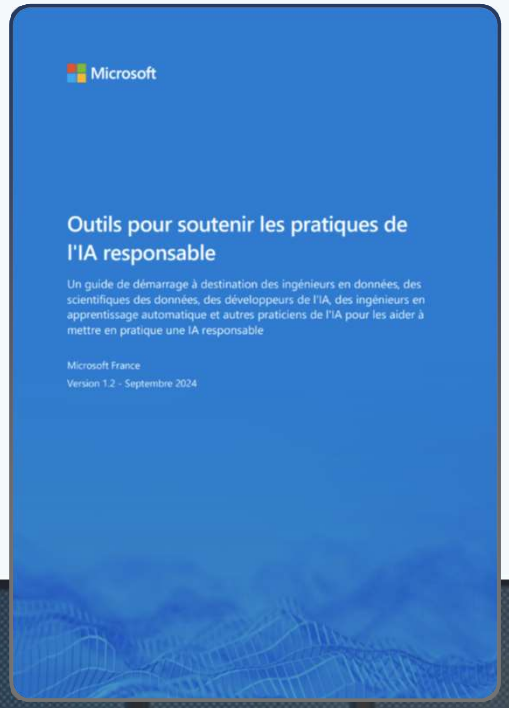
  to help improve the AI system

Looking ahead

1. Innovative new approaches to responsible AI development in our own products.
2. Creating tools for our customers to responsibly develop their own (generative) AI systems.
3. Sharing our learnings and best practices with the responsible AI ecosystem at large.
4. Supporting the development of laws, norms, and standards via broad and inclusive multistakeholder processes.

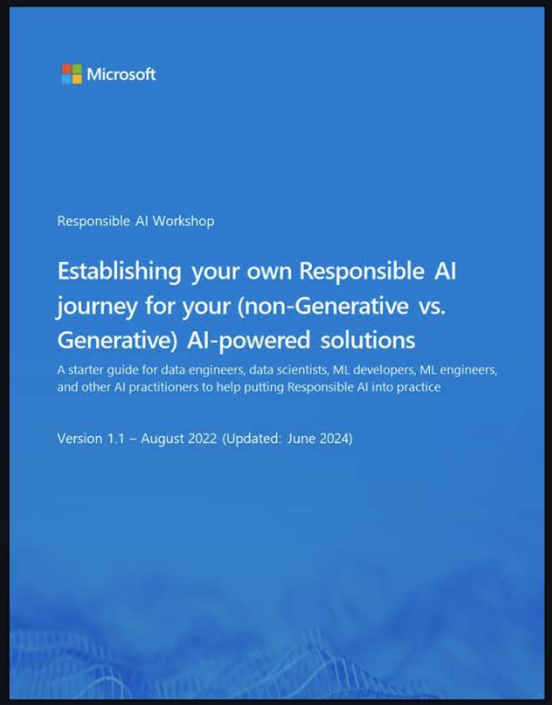


Pour (bien) démarrer, visiter le site :
https://aka.ms/**RAIresources**



Français - Anglais

github.com/microsoft/responsible-ai-workshop



Microsoft

Merci !