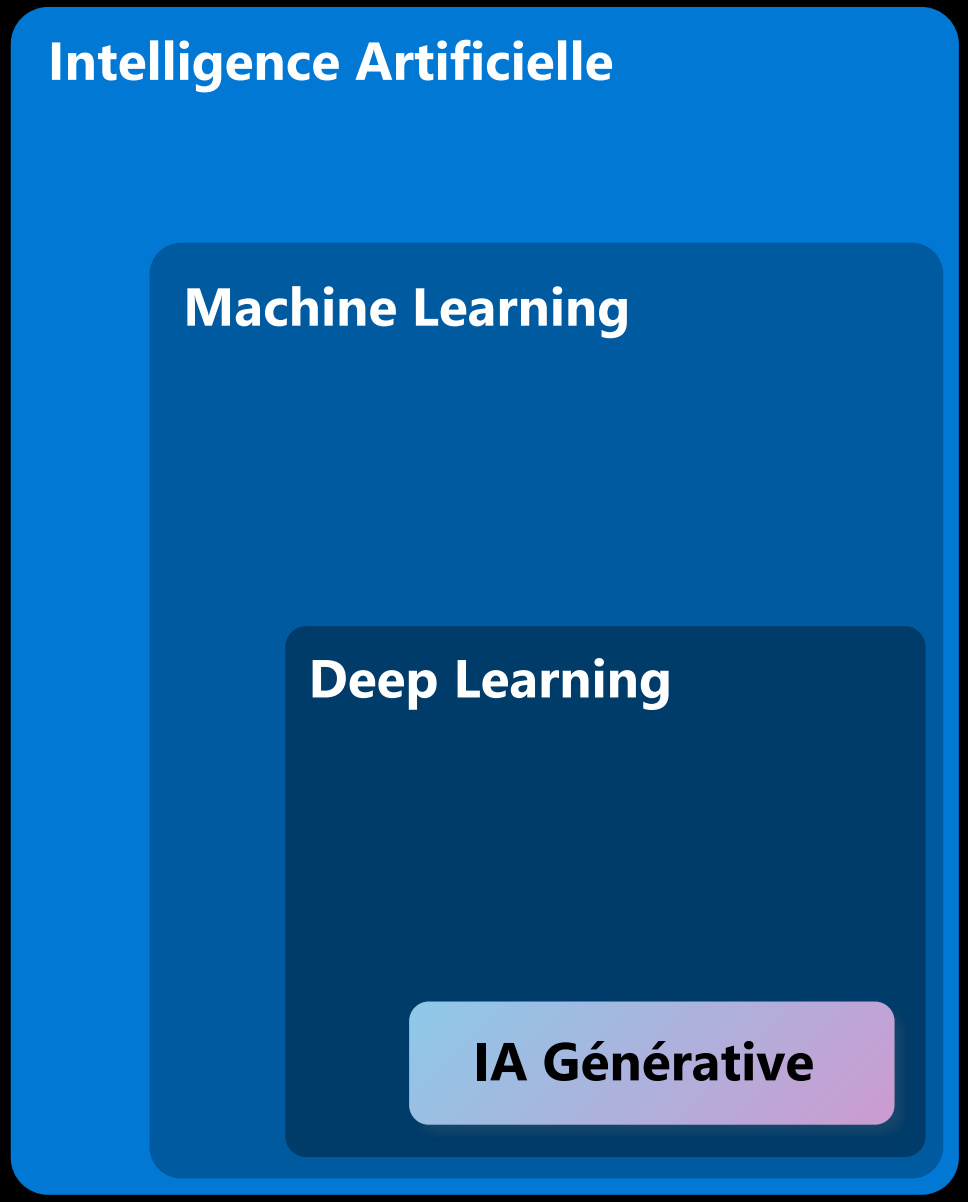
ASSOCIATION
EUROPE
FINANCES
RÉGULATIONS

L'IA, composantes techniques

Marc Gardette

Deputy CTO, Microsoft France

La trajectoire de l'IA



1956

Intelligence Artificielle

domaine de l'informatique qui cherche à créer des machines intelligentes capables de reproduire ou de dépasser l'intelligence humaine

90's

Machine Learning

sous-ensemble de l'IA qui permet aux machines d'apprendre à partir de données existantes et d'améliorer ces données pour prendre des décisions ou faire des prédictions

20's

Deep Learning

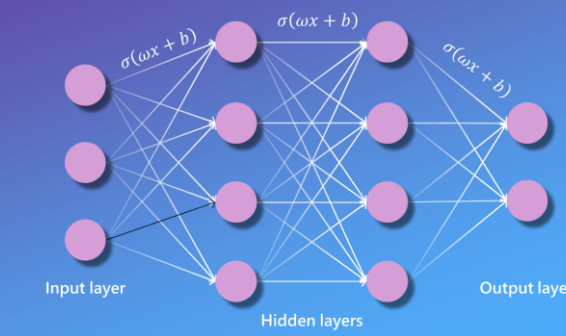
Technique d'apprentissage automatique dans laquelle des couches de réseaux neuronaux sont utilisées pour traiter des données et prendre des décisions

2021

IA Générative

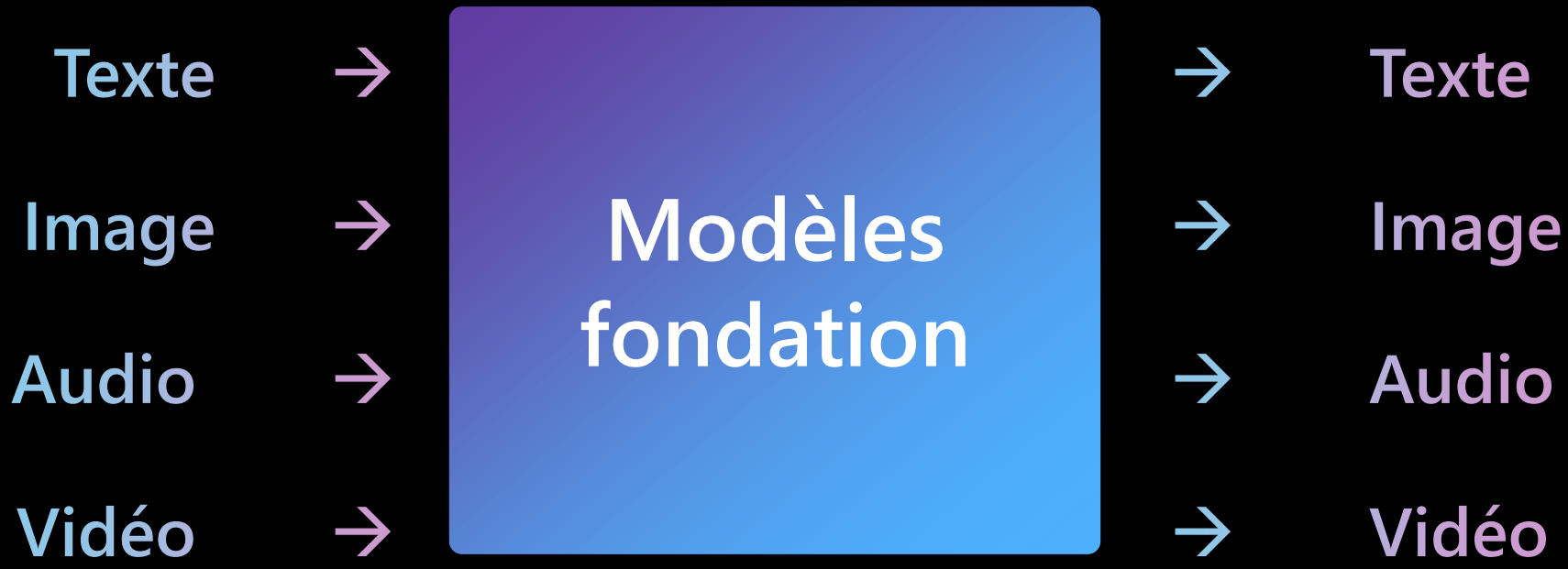
Créer de nouveaux contenus écrits, visuels et auditifs à partir d'invites ou de données existantes

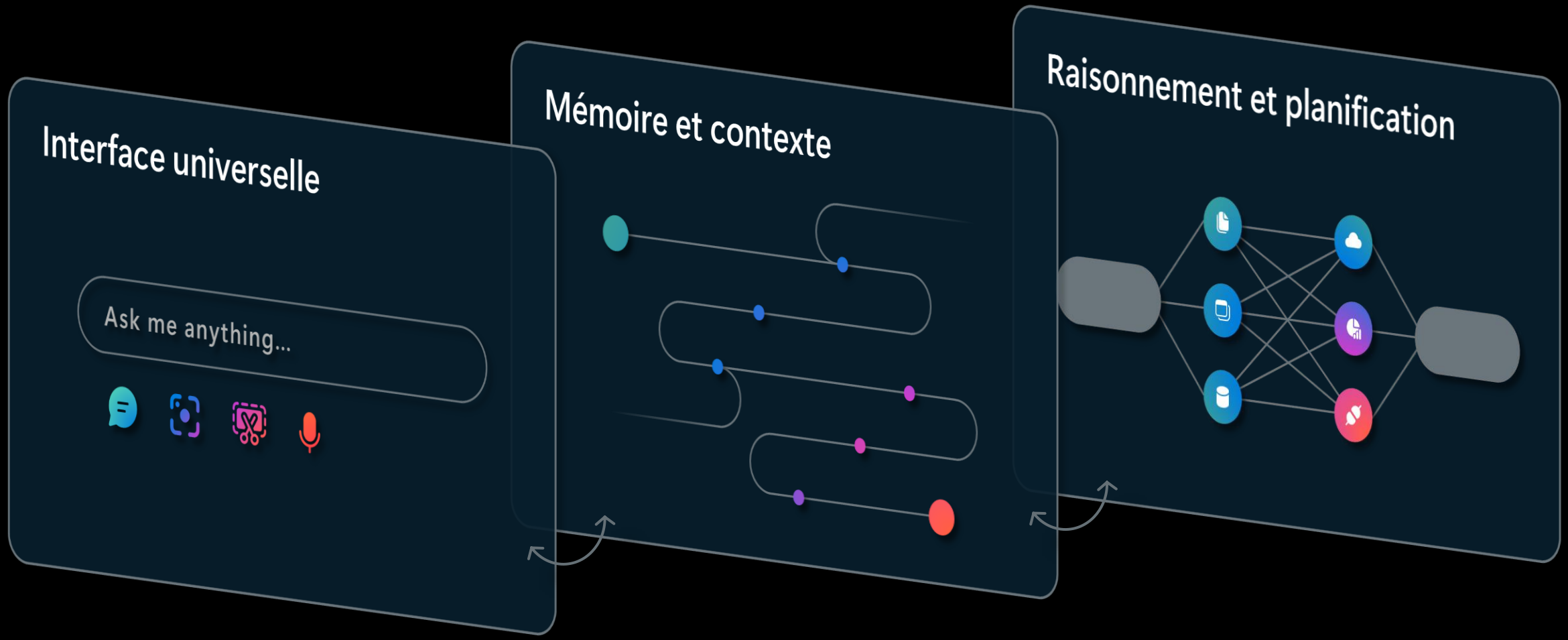
Large Language Model

Texte (Prompt) → LLM  → Texte (Réponse)

Propriétés émergentes

Modèle fondation multimodal




Interface universelle
Ask me anything...
Mémoire et contexte
Raisonnement et planification

Trois indicateurs pour mesurer la valeur

VITESSE

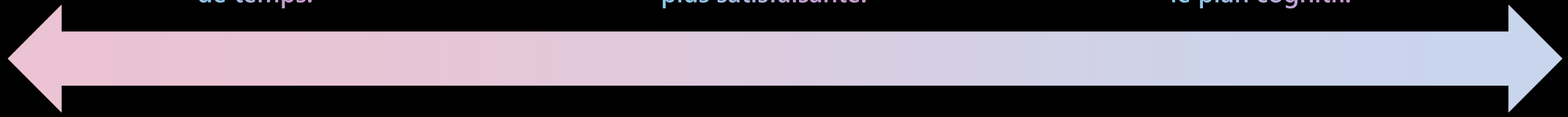
Je peux trouver des informations ou accomplir des tâches en moins de temps.

QUALITÉ

Je peux obtenir de meilleurs résultats ou une expérience plus satisfaisante.

PROFONDEUR

Je peux découvrir et accomplir des tâches complexes et exigeantes sur le plan cognitif.



Tâches répétitives ou laborieuses qui se prêtent à l'automatisation

Des raisonnements plus complexes nécessitant une solide expertise

Une grande diversité de modèles

Azure OpenAI Service

GPT-4o

GPT-4-Turbo with Vision, GPT-4, GPT-3.5

Embeddings

DALL·E

Whisper, Text to speech

Phi models

Phi-3-mini

Phi-3-small

Phi-3-medium

Phi-3-vision

Meta

Llama-2-70b/70b-chat

Llama-2-13b/13b-chat

Llama-2-7b/7b-chat

Llama-3

CodeLlama

Mistral AI

Mistral Large

Mistral 7B

Mixtral 8x7B – Mixture of Experts

Cohere

Cohere R+

Cohere R

Embed v3-Multilingual

Embed v3-English

Hugging Face

Falcon/TII

Stable Diffusion/Stability AI

Dolly/Databricks

CLIP/OpenAI

Databricks

Databricks/dbrx-base

Databricks/dbrx-instruct

NVIDIA

Nemotron-3-8B-4k

Nemotron-3-8B-Chat-SFT/RLHF/ SteerLM

Nemotron-3-8B-QA

Snowflake

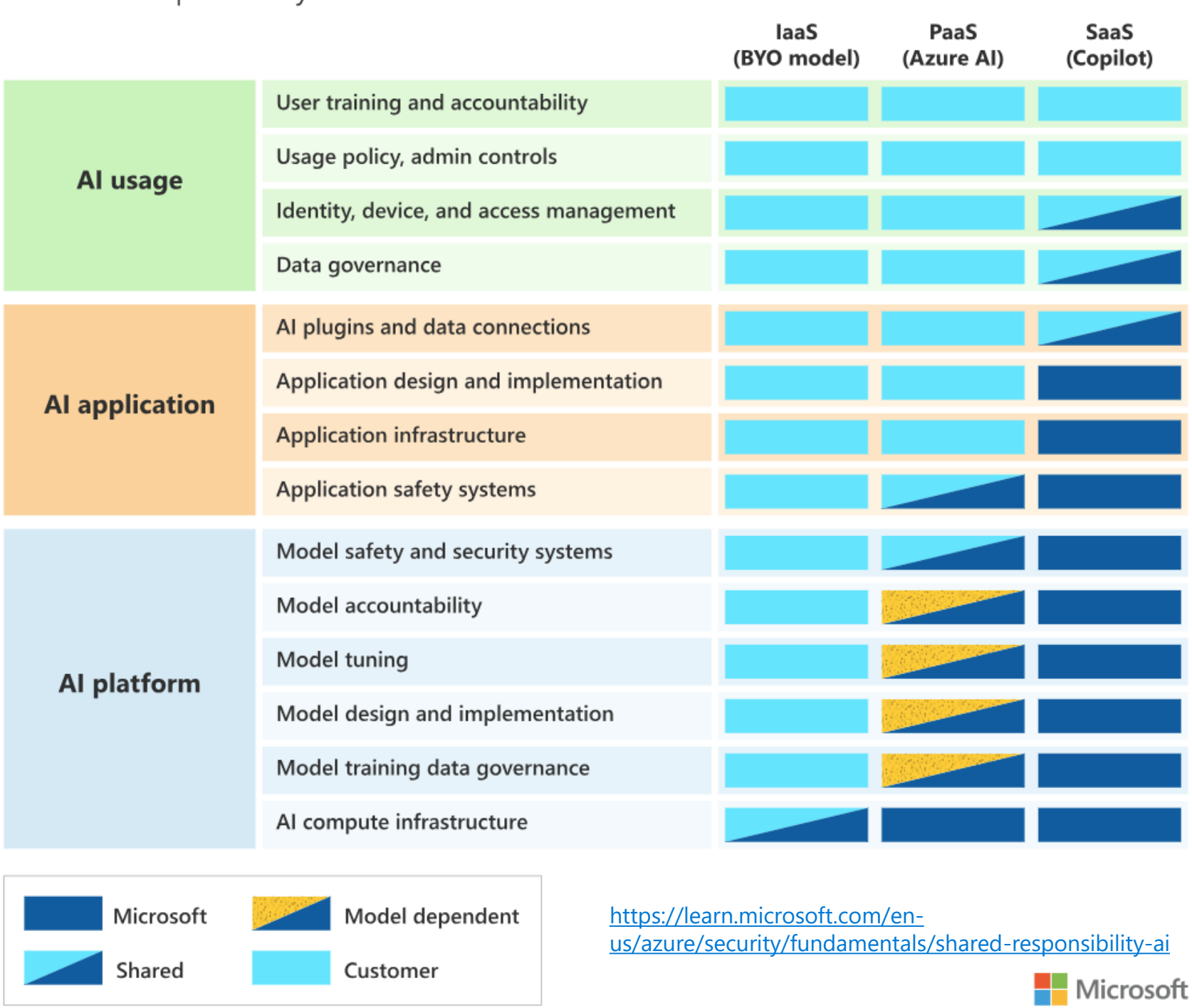
Snowflake/arctic-base

Snowflake/arctic-instruct

Nouvelle économie de l'IA

[Microsoft's AI Access Principles: Our commitments to promote innovation and competition in the new AI economy]

Modèle de responsabilité partagée de l'IA



https://learn.microsoft.com/en-us/azure/security/fundamentals/shared-responsibility-ai

Microsoft

MITRE ATT&CK

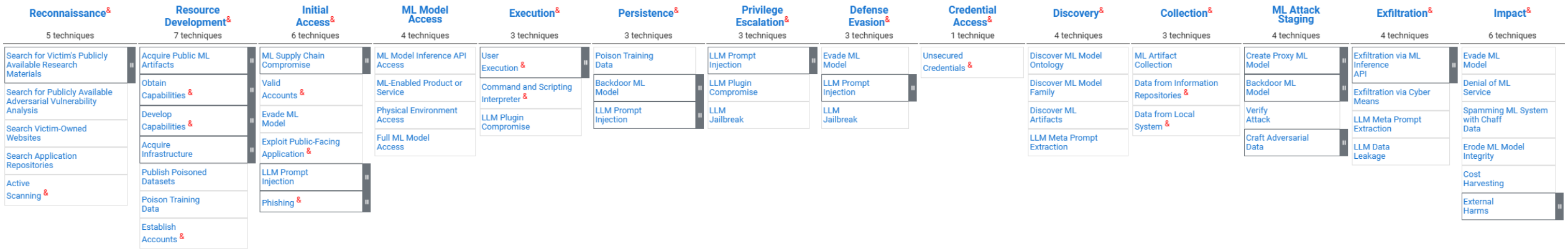
MITRE ATLAS™

Adversarial Threat Landscape for Artificial-Intelligence Systems

Une base de connaissances interactive de la communauté au format MITRE ATT&CK® à exploiter pour aider à protéger l'IA et les applications d'IA générative : tactiques, techniques et procédures

ATLAS Matrix

The ATLAS Matrix below shows the progression of tactics used in attacks as columns from left to right, with ML techniques belonging to each tactic below. & indicates an adaption from ATT&CK. Click on the blue links to learn more about each item, or search and view ATLAS tactics and techniques using the links at the top navigation bar. View the ATLAS matrix highlighted alongside ATT&CK Enterprise techniques on the ATLAS Navigator.



MITRE and Microsoft Collaborate to Address Generative AI Security Risks Microsoft and MITRE Create Tool to Help Security Teams Prepare for Attacks on Machine Learning Systems

IA générative : un nouvel ensemble de risques

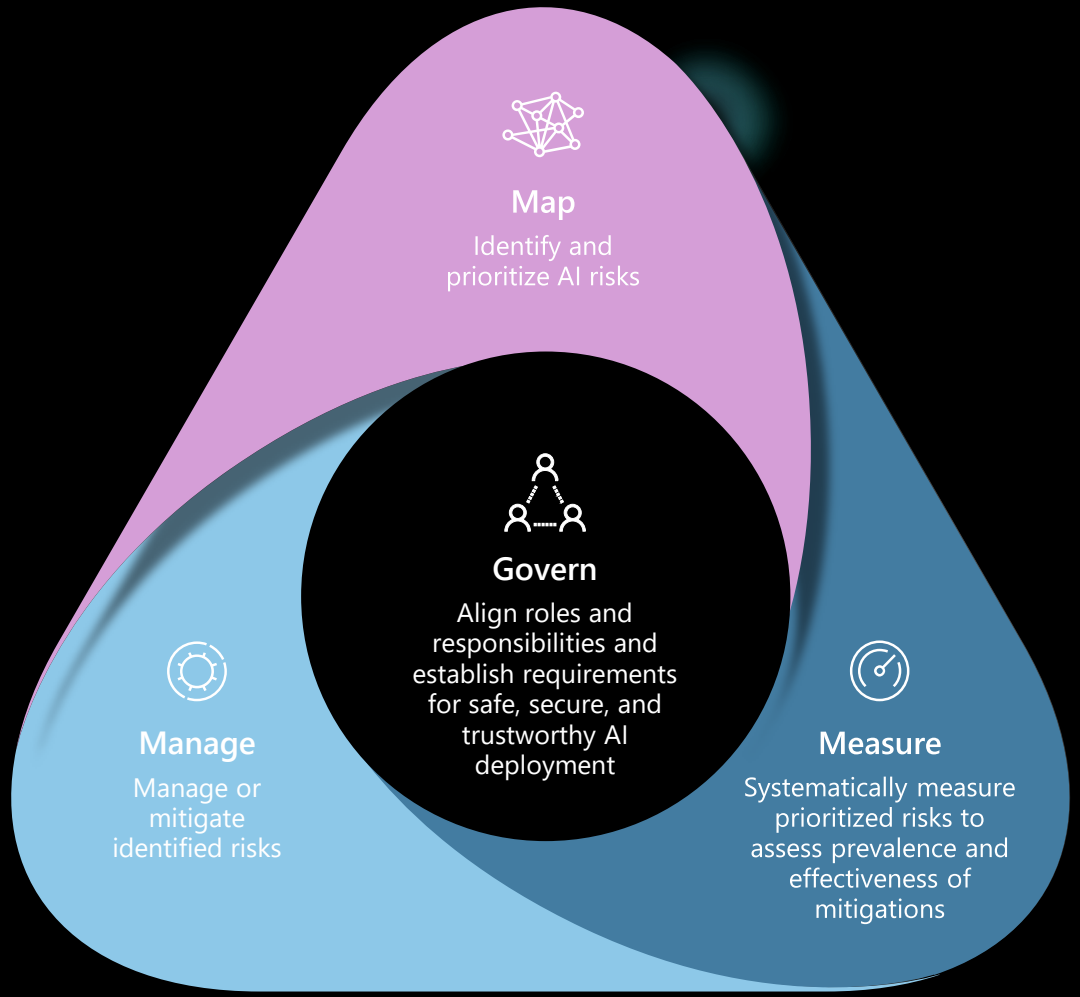
Risques liés à l'IA responsable

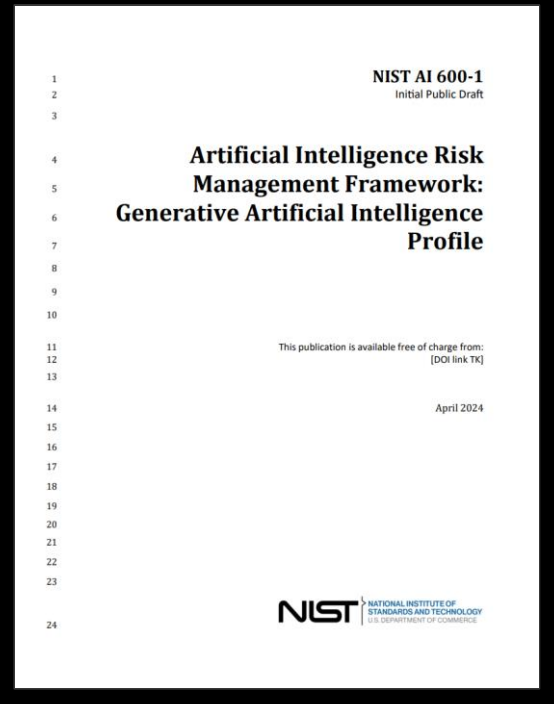
- Contenu non fondé
- Contenu nuisible, offensant, violent
- Bias
- Désinformation et propagande.
- Usurpation d'identité humaine
- Propriété intellectuelle
- Vie privée et confidentialité des données

Risques de sécurité

- Jail breaking
- Indirect prompt injection
- Attaques par empoisonnement de donnée
- Model backdoor
- Vol de modèle
- Exfiltration de données
- DDOS / Wallet (GPU abuse)
- Exécution de code à distance via les plugins

Développer des des applications d'IA générative de manière responsable





NIST AI 600-1
Initial Public Draft

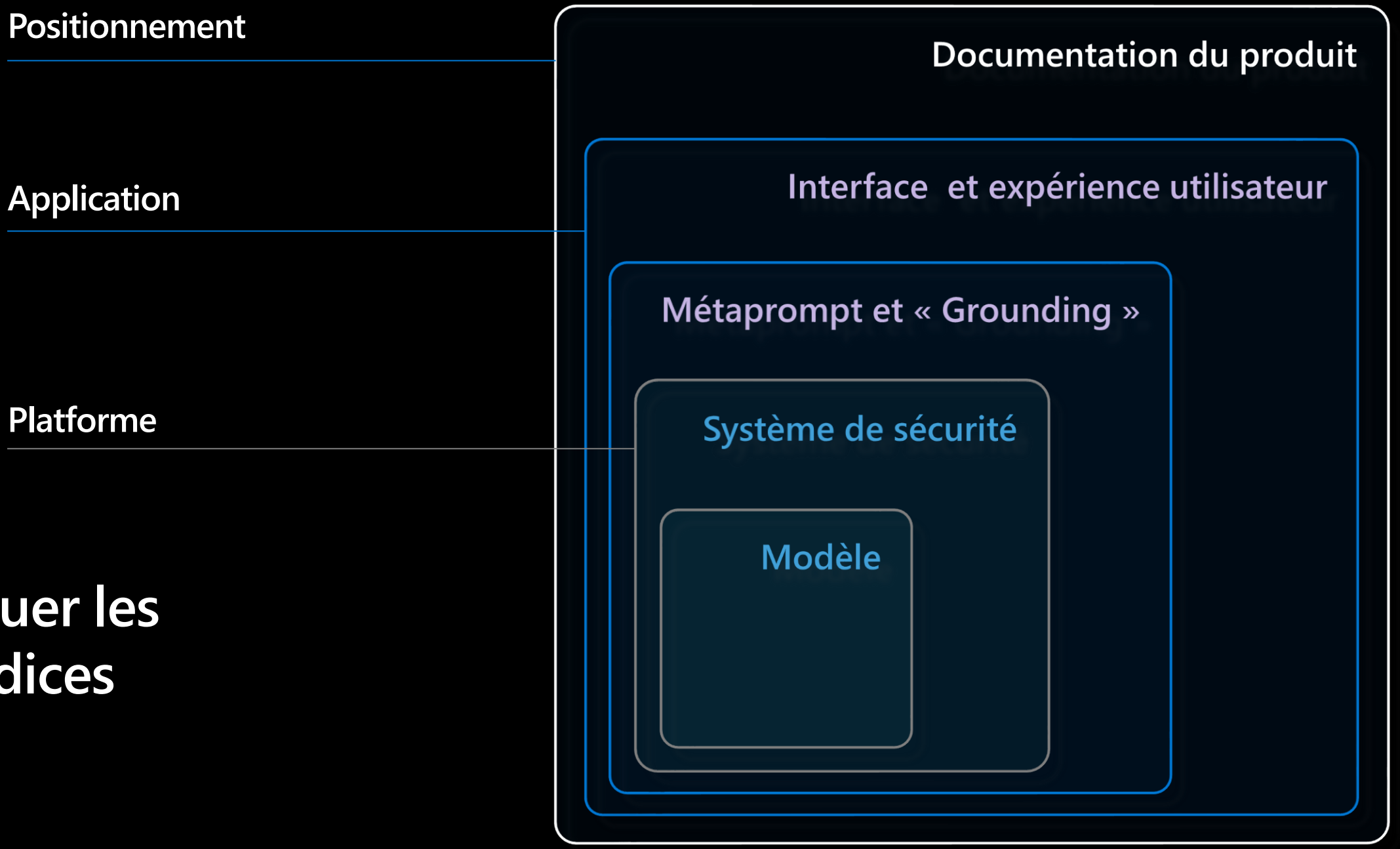
Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile

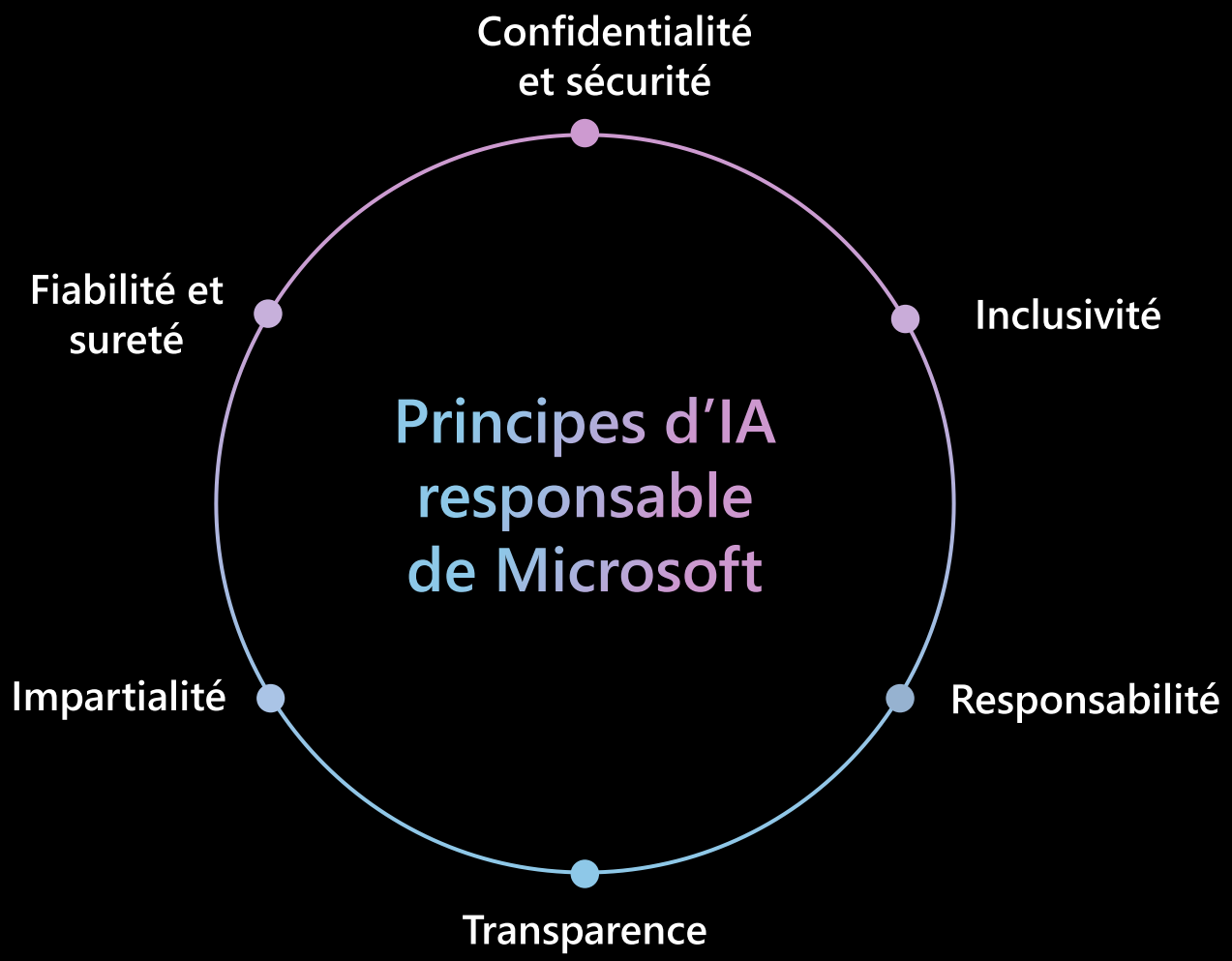
This publication is available free of charge from:
[DOI link TK]

April 2024

NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY U.S. DEPARTMENT OF COMMERCE

Atténuer les préjudices





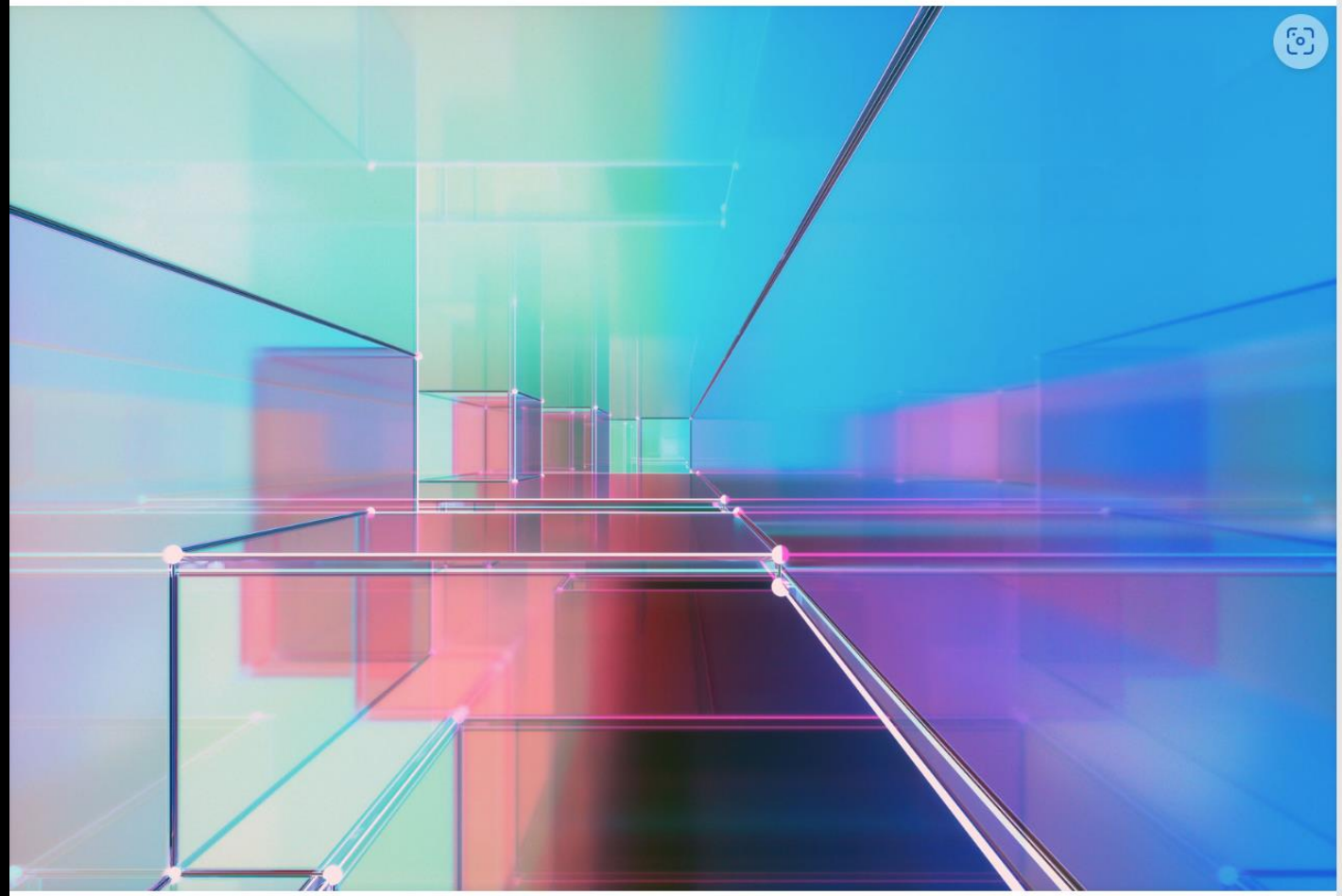
Les éléments constitutifs de la mise en œuvre des principes

- Outils et processus
- Formation et pratiques
- Règles
- Gouvernance

Responsible AI Transparency report

Providing further transparency on our responsible AI efforts

May 1, 2024 | Brad Smith, Vice Chair & President; Natasha Crampton, Chief Responsible AI Officer



The following is the foreword to the inaugural edition of our annual Responsible AI Transparency Report. The FULL REPORT is available at this link.

- ANNEXE

  - What Is an AI Anyway? | Mustafa Suleyman | TED

  When it comes to artificial intelligence, what are we actually creating? Even those closest to its development are struggling to describe exactly where things are headed, says Microsoft AI CEO Mustafa Suleyman, one of the primary architects of the AI models many of us use today. He offers an honest and compelling new vision for the future of AI, proposing an unignorable metaphor — a new digital species — to focus attention on this extraordinary moment.

  - What Is an AI Anyway? | Mustafa Suleyman | TED (youtube.com)

  Inside AI Security with Mark Russinovich

  Join Mark Russinovich to explore the landscape of AI security, focusing on threat modeling, defense tactics, our red teaming approaches, and the path to confidential AI. You will learn about various kinds of attacks in AI systems and our defenses, such as backdoors, poison data, prompt injection attacks, and more.

  Inside AI Security with Mark Russinovich | BRK227 (youtube.com)