

L'IA, composantes techniques

Marc Gardette

Deputy CTO, Microsoft France



La trajectoire de l'IA

Intelligence Artificielle

Machine Learning

Deep Learning

IA Générative



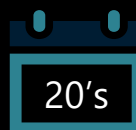
Intelligence Artificielle

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



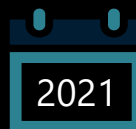
Machine Learning

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



Deep Learning

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



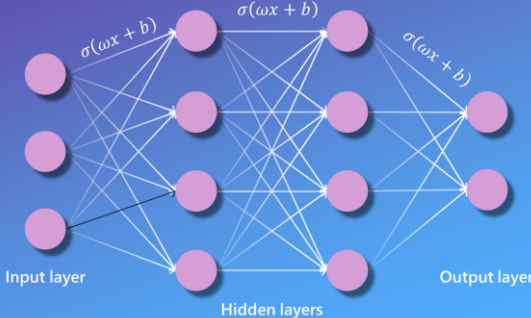
IA Générative

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

Large Language Model

Texte (Prompt) →

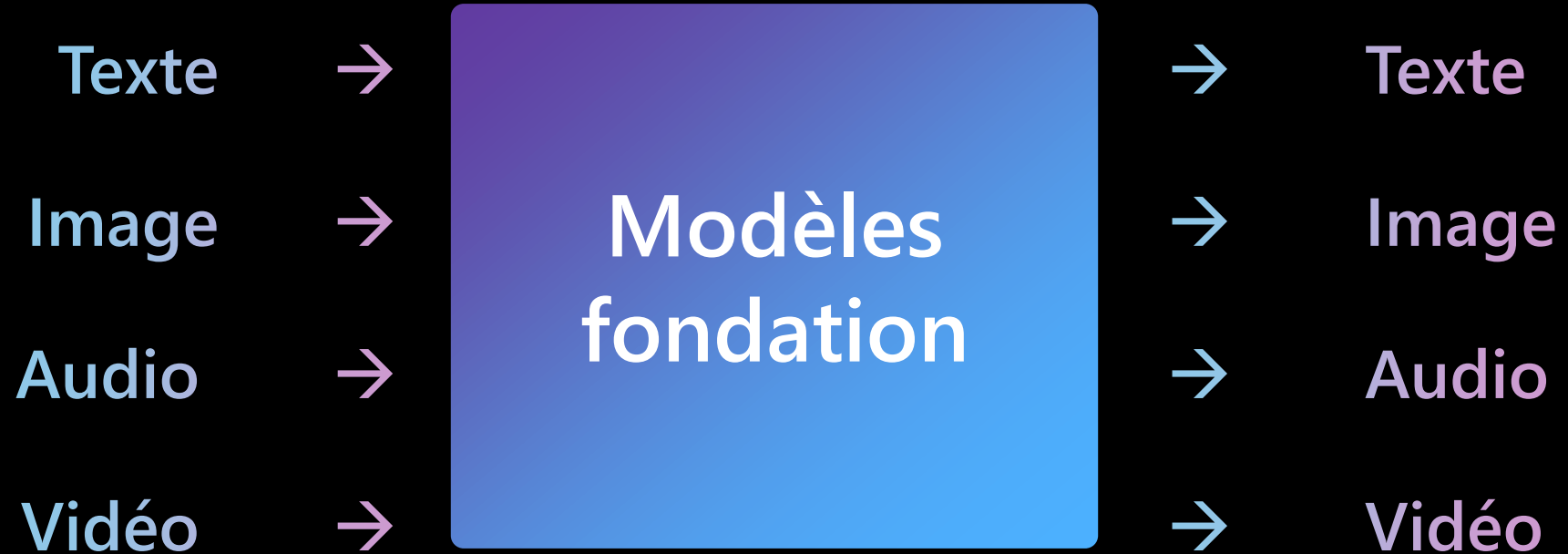
LLM



→ Texte (Réponse)

Propriétés émergentes

Modèle fondation multimodal

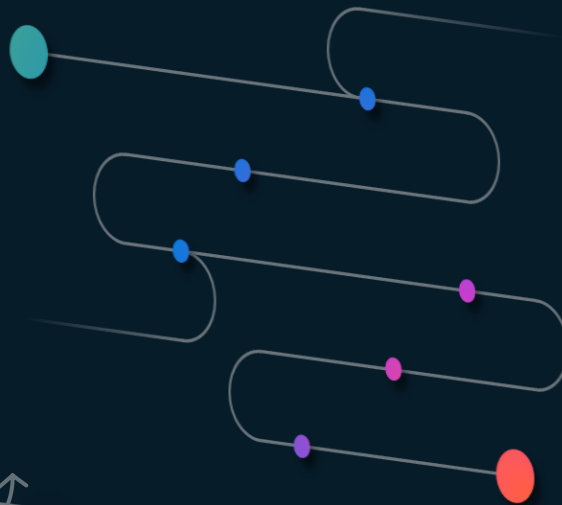


Interface universelle

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

Chips



**AI datacenters
(infrastructure)**



Data



Foundation models



Tooling



Applications



Distribution



**Developers
and
users**



Electrical power + connectivity

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

Modèle de responsabilité partagée de l'IA

		IaaS (BYO model)	PaaS (Azure AI)	SaaS (Copilot)
AI usage	User training and accountability	Customer	Customer	Customer
	Usage policy, admin controls	Customer	Customer	Customer
	Identity, device, and access management	Customer	Customer	Shared
	Data governance	Customer	Customer	Shared
AI application	AI plugins and data connections	Customer	Customer	Shared
	Application design and implementation	Customer	Customer	Microsoft
	Application infrastructure	Customer	Customer	Microsoft
	Application safety systems	Customer	Shared	Microsoft
AI platform	Model safety and security systems	Customer	Shared	Microsoft
	Model accountability	Customer	Shared	Microsoft
	Model tuning	Customer	Shared	Microsoft
	Model design and implementation	Customer	Shared	Microsoft
	Model training data governance	Customer	Shared	Microsoft
	AI compute infrastructure	Shared	Microsoft	Microsoft



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

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](#).

Reconnaissance &	Resource Development &	Initial Access &	ML Model Access	Execution &	Persistence &	Privilege Escalation &	Defense Evasion &	Credential Access &	Discovery &	Collection &	ML Attack Staging	Exfiltration &	Impact &
5 techniques	7 techniques	6 techniques	4 techniques	3 techniques	3 techniques	3 techniques	3 techniques	1 technique	4 techniques	3 techniques	4 techniques	4 techniques	6 techniques
Search for Victim's Publicly Available Research Materials	Acquire Public ML Artifacts	ML Supply Chain Compromise	ML Model Inference API Access	User Execution &	Poison Training Data	LLM Prompt Injection	Evade ML Model	Unsecured Credentials &	Discover ML Model Ontology	ML Artifact Collection	Create Proxy ML Model	Exfiltration via ML Inference API	Evade ML Model
Search for Publicly Available Adversarial Vulnerability Analysis	Obtain Capabilities &	Valid Accounts &	ML-Enabled Product or Service	Command and Scripting Interpreter &	Backdoor ML Model	LLM Plugin Compromise	LLM Prompt Injection		Discover ML Model Family	Data from Information Repositories &	Backdoor ML Model	Exfiltration via Cyber Means	Denial of ML Service
Search Victim-Owned Websites	Develop Capabilities &	Evade ML Model	Physical Environment Access	LLM Plugin Compromise	LLM Prompt Injection	LLM Jailbreak	LLM Jailbreak		Discover ML Artifacts	Data from Local System &	Verify Attack	LLM Meta Prompt Extraction	Spamming ML System with Chaff Data
Search Application Repositories	Acquire Infrastructure	Exploit Public-Facing Application &	Full ML Model Access						LLM Meta Prompt Extraction		Craft Adversarial Data	LLM Data Leakage	Erode ML Model Integrity
Active Scanning &	Publish Poisoned Datasets	LLM Prompt Injection											Cost Harvesting
	Poison Training Data	Phishing &											External Harms
	Establish Accounts &												

[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

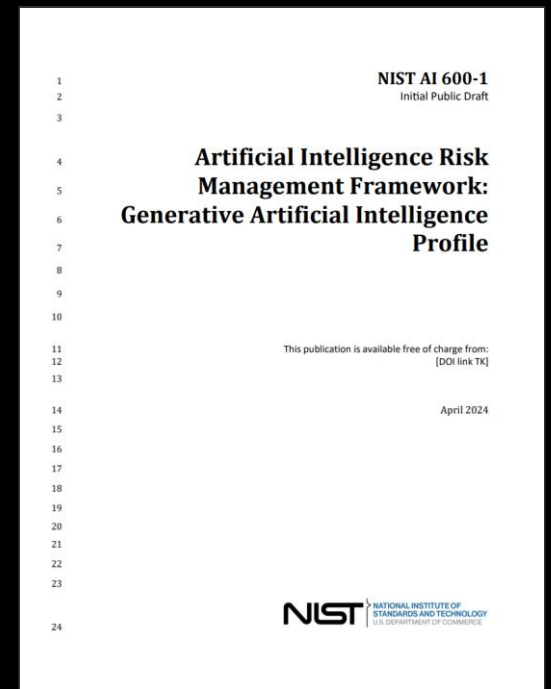
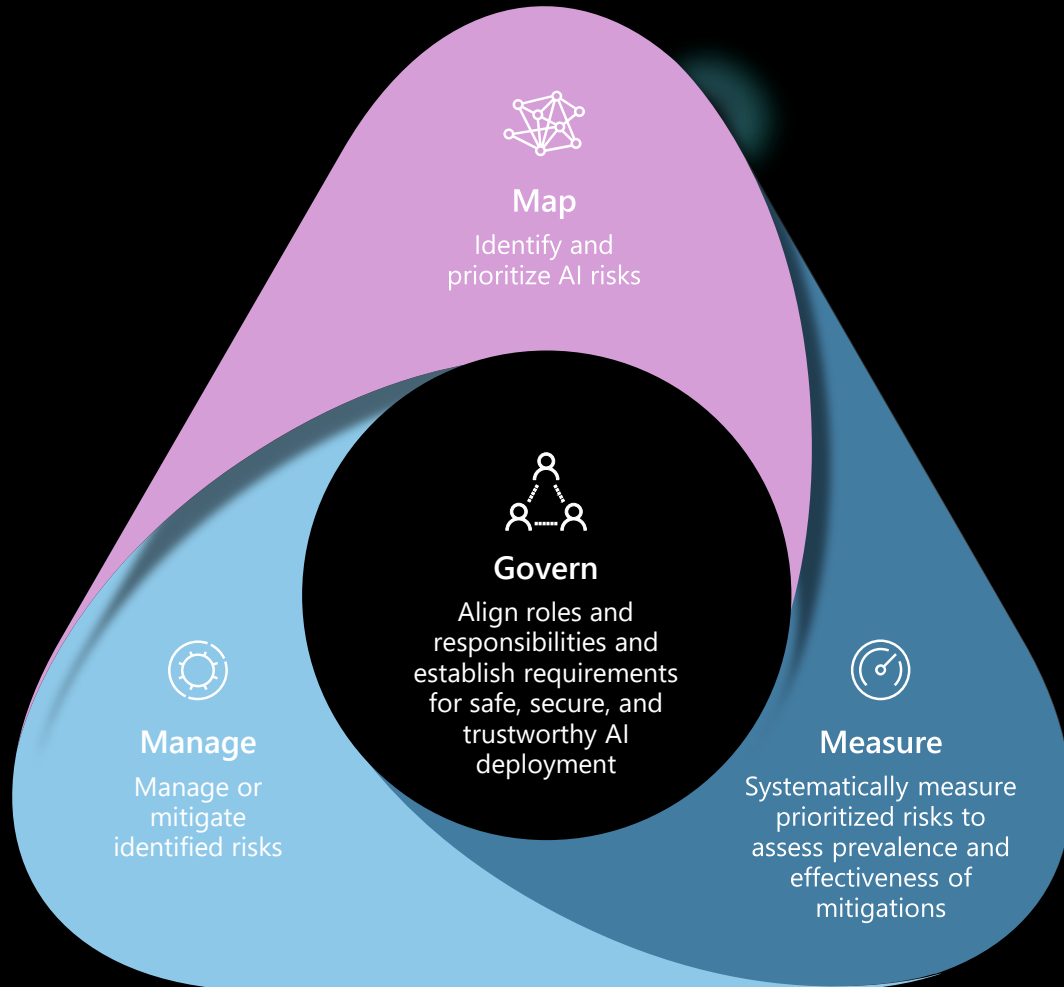
Risques liés à l'IA responsable

- 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



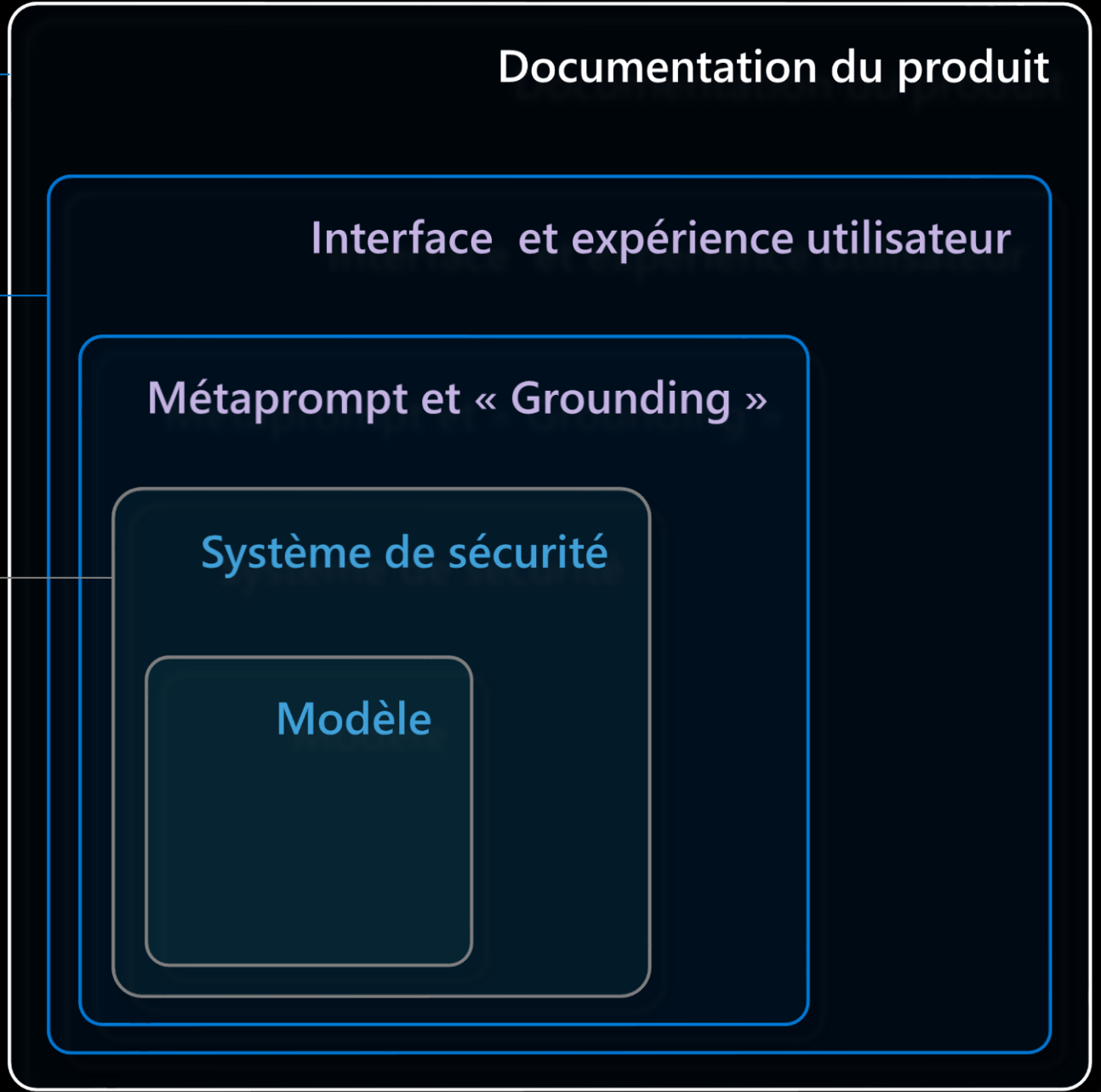
4	2. Overview of Risks Unique to or Exacerbated by GAI	3
5	2.1. CBRN Information	4
6	2.2. Confabulation	5
7	2.3. Dangerous or Violent Recommendations	5
8	2.4. Data Privacy	6
9	2.5. Environmental	6
10	2.6. Human-AI Configuration	7
11	2.7. Information Integrity	7
12	2.8. Information Security	8
13	2.9. Intellectual Property	8
14	2.10. Obscene, Degrading, and/or Abusive Content	9
15	2.11. Toxicity, Bias, and Homogenization	9
16	2.12. Value Chain and Component Integration	10
17	3. Actions to Manage GAI Risks	11

Positionnement

Application

Plateforme

Atténuer les
préjudices



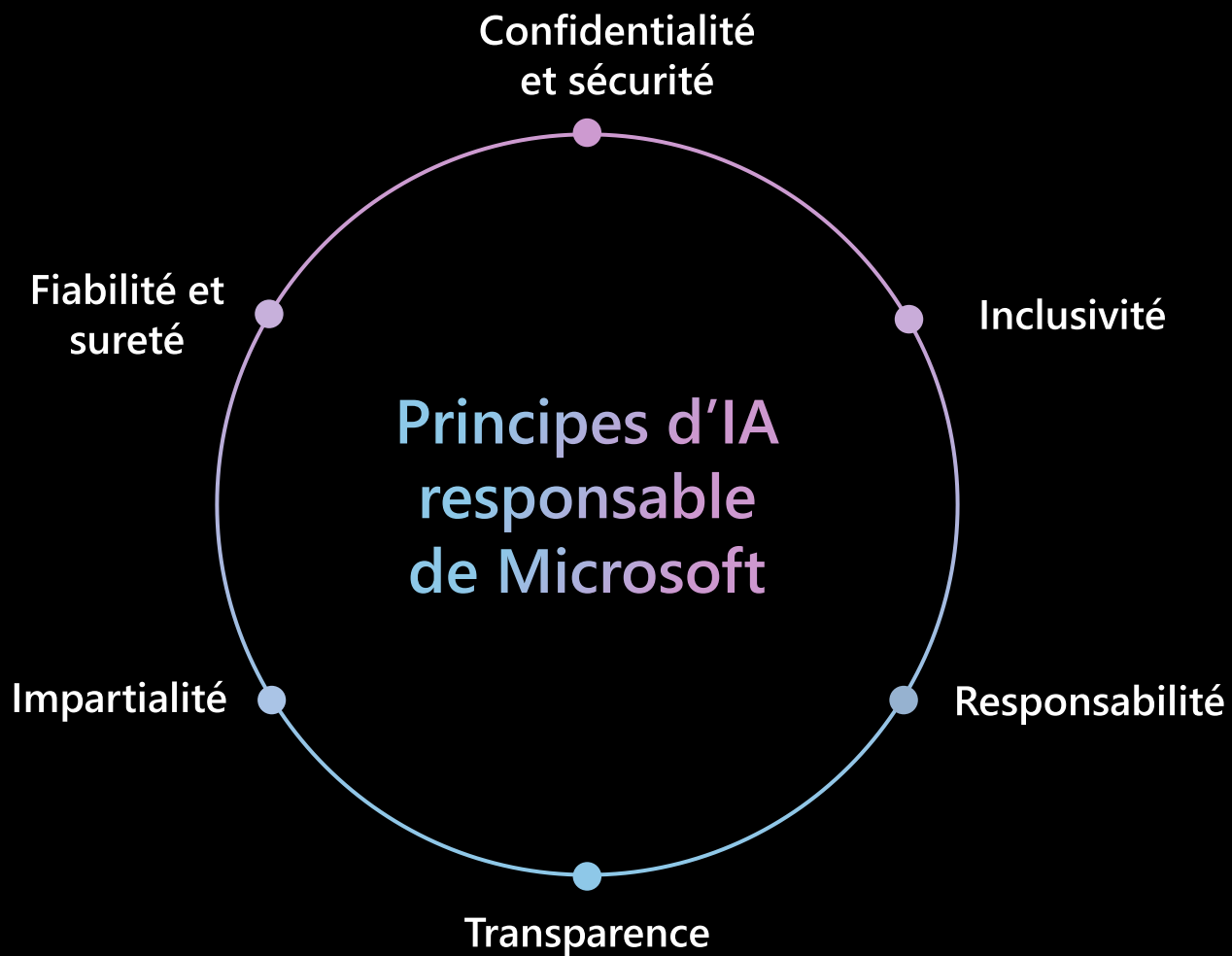
Documentation du produit

Interface et expérience utilisateur

Métaprompt et « Grounding »

Système de sécurité

Modèle



Les éléments constitutifs de la mise en œuvre des principes



Outils et processus



Formation et pratiques



Règles



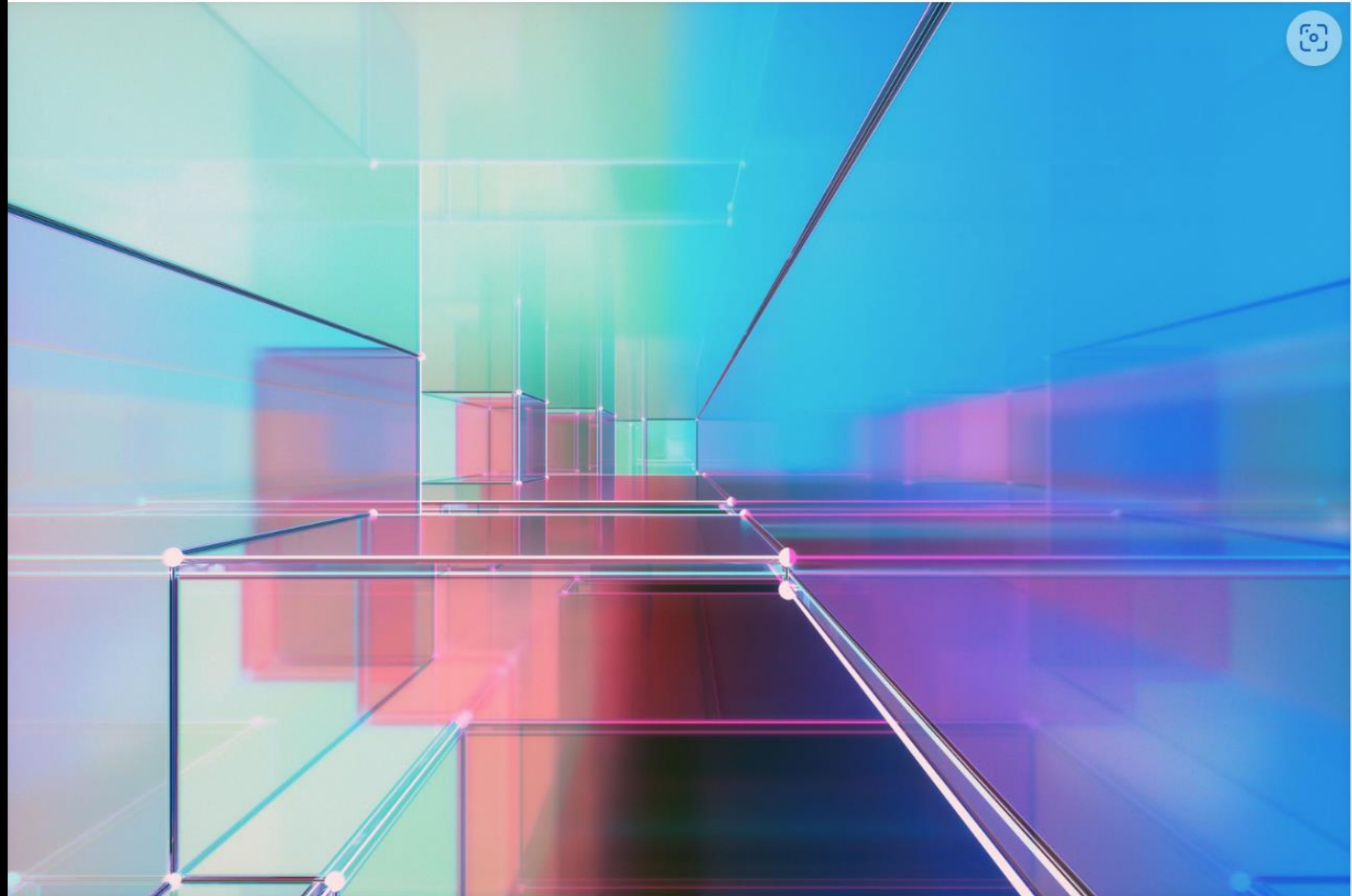
Gouvernance

Providing further transparency on our responsible AI efforts

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



Responsible AI Transparency report



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 \(youtube.com\)](#)