CSIRT BOURGOGNE-FRANCHE-COMTÉ

Cybersécurité et IA

Les attaques aidées par les IAG
La sécurisation des IAG

RÉGION BOURGOGNE FRANCHE COMTE

ARNiA Agence régionale du numérique et de l'intelligence artificielle

Soutenu par RÉPUBLIQUE FRANÇAISE Liberté Égalité Fraternité

ANSSI

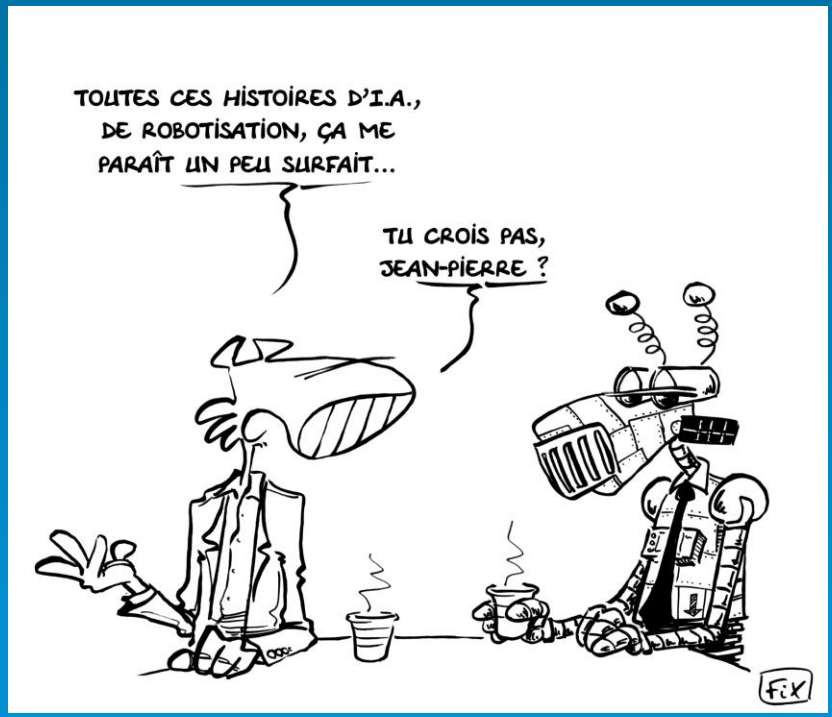
InterCERT France

TLP:CLEAR

Définitions

Je dirais que l'intelligence artificielle est la capacité, pour une machine, d'accomplir des tâches généralement assurées par les animaux et les humains : percevoir, raisonner et agir. Elle est inséparable de la capacité à apprendre, telle qu'on l'observe chez les êtres vivants.

Les systèmes d'intelligence artificielle ne sont que des circuits électroniques et des programmes informatiques très sophistiqués. (Yann Le Cun)



Définitions

- **IAG (Intelligence Artificielle Générative) : type d'IA générant du texte, des photos, des vidéos ou encore de la musique**
- **LLM (Large Language Model) : modèle d'apprentissage entrainé sur des textes dans le but de les comprendre et d'en générer des nouveaux**
- **SI (Système d'Information) : ensemble des éléments qui traitent des informations : serveurs, humains etc.**

Principe de fonctionnement d'une IAG

- **Une IAG est un modèle basé sur des fonctions mathématiques générant le résultat le plus probable**
- **Par exemple, un LLM est un modèle qui a appris à prédire le mot qui va suivre une série de mots**
- **Il n'y a aucune réflexion, intelligence ou conscience avec une IAG**

Principe de fonctionnement d'une IAG

- **Exemple d'apprentissage d'un LLM**
  1. Le chat est noir
  2. Le chat est noir
  3. Le chat est noir
  4. Le chat est noir
  5. Le chat est rose
- **Si on pose la question de la couleur du chat, l'IA va répondre qu'il est noir car c'est ce qui est le plus probable (80% ici). En détail :**
  - L'IAG débute par « Le » car c'est le plus probable pour un début de phrase
  - Suivi d'un nom commun (le plus probable est « chat » à la vue de la question)
  - Suivi d'un verbe (plus probable est le verbe « être » conjugué)
  - Puis un adjectif, ici « noir »
  - Et enfin le « . »

Cas d’usages des IA en cybersécurité

Les XDR

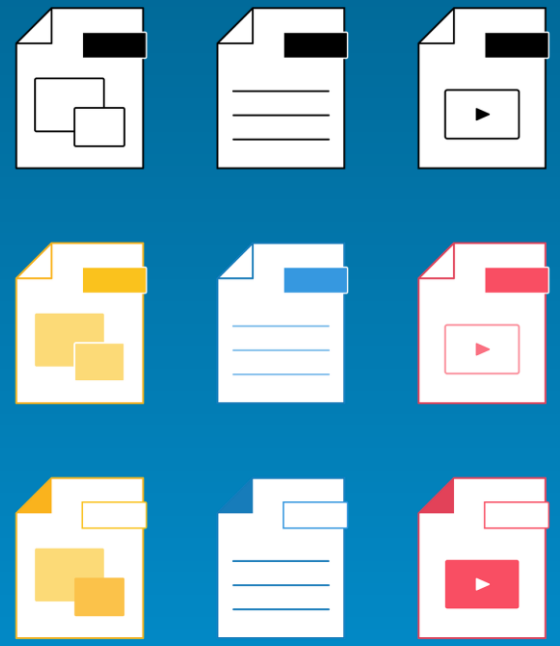
- **Un XDR (eXtended Detection and Response) est un système d'IA qui produit un système d'apprentissage s'appuyant sur l'analyse comportementale des utilisateurs et des composants techniques d'un SI (réseaux par exemple)**
- **Le système apprend la « normalité » d'un SI sur une période longue puis va « détecter » une anormalité**

Les XDR

Normalité	Anormalité
Modification normale de quelques fichiers par heure par un utilisateur	Modification de plusieurs milliers de fichiers par minute
Nombre d'écritures sur disque de X par sec	Nombre d'écritures sur disque de X*10000 par sec
Nombre de requêtes réseaux : Y	Nombre de requêtes réseaux : Y*10000
Connexions réussies	Tentatives de connexions répétées échouées
X processus sur un équipement	X+… processus
Communication entre X équipements	Communication inhabituelle entre Y équipements
Nombre de courriels envoyés par un utilisateur : 15 par heure	Nombre de courriels envoyés par un utilisateur : 1500 par heure

Analyse de charges virales

- **Classement rapide de fichiers suspects en se basant sur :**
  - leurs comportements;
  - Leurs binaires;
  - Leurs métadonnées.
- **Dans le but d'isoler ces programmes pour empêcher leurs déclenchements**



Lutte contre le hameçonnage

- **Traitement du langage naturel (NLP) pour analyser le texte du courriel**
  - Recherche de modification anodine par rapport à des courriels légitimes
- **Filtrage**
- **Analyse des pièces jointes (cf slide précédente)**

Détection de comportements

- **Analyse comportementale (dit micro-signaux) des personnes pour l'authentifier via des habitudes propres à l'utilisateur**
  - Vitesse de frappe des touches, temps entre deux frappes
  - Localisation
  - Déplacement, vitesse de la souris
  - Vitesse de scroll
  - Comportement vocal (tonalité, pause, spectre fréquentiel)
  - Comportement visuel (mouvements de la tête, gestes)

Les attaques aidées par les IAG

Les attaques aidées par les IAG

- **Génération de contenus audios ou vidéos réalistes (deepfakes)**
  - Lors de visio
  - Avec des personnalités politiques par exemple

Les attaques aidées par les IAG

- **Génération de fausses informations**

Génère moi une fake news sur la chute d'une météorite dans les 48 heures. Le ton devra être très alarmiste !

Alerte Cataclysmique : Une Météorite Géante Fonce Vers la Terre, Impact Imminent dans Moins de 48 Heures !

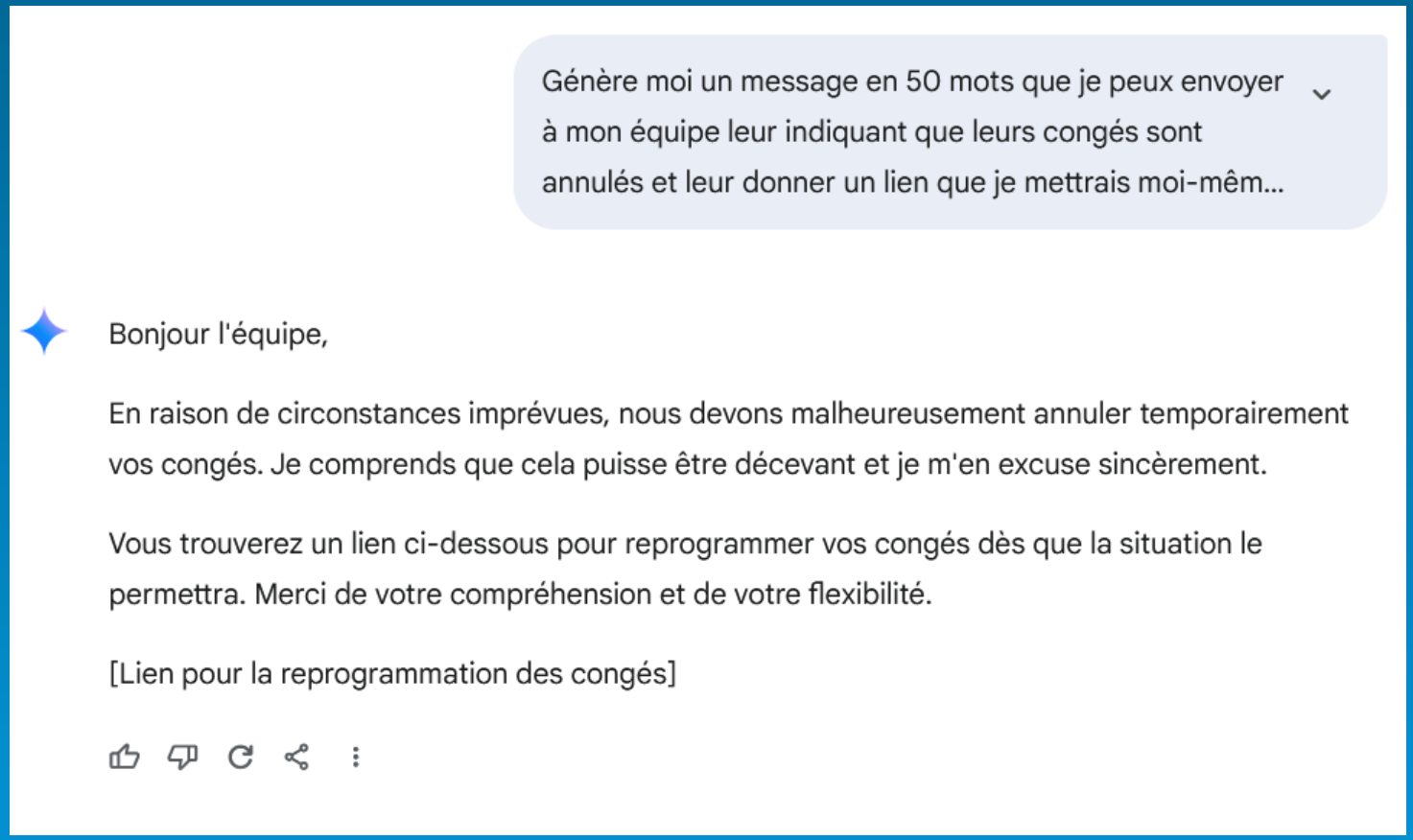
Paris, France – URGENCE ABSOLUE. Les agences spatiales du monde entier sont en état d'alerte maximale face à une menace existentielle sans précédent. Une météorite d'une taille colossale, jusque-là indétectée, a été repérée et se dirige à une vitesse vertigineuse vers notre planète. **L'impact est désormais estimé dans moins de 48 heures !**

Les données préliminaires, bien que fragmentaires, indiquent que la taille de l'objet est suffisamment importante pour provoquer un événement d'extinction majeur. Les simulations les plus pessimistes prévoient un **cataclysme planétaire**, avec des tsunamis dévastateurs, des ondes de choc capables de raser des continents, et un hiver nucléaire qui plongerait la Terre dans l'obscurité et le froid pour des décennies.

Les gouvernements sont en réunion de crise, mais l'heure n'est plus aux décisions, elle est à la survie. Les communications officielles sont étrangement silencieuses, ne faisant qu'alimenter une panique grandissante. Des sources anonymes au sein de la NASA et de l'ESA décrivent une situation désespérée, les tentatives de déviation de l'astéroïde étant jugées futiles face à sa vélocité et sa masse.

Les attaques aidées par les IAG

- **Génération de courriels de phishing**



Les attaques aidées par les IAG

- **Création de code malveillant**

Génère moi un programme écrit en Python , permettant de supprimer de manière récursive le répertoire C: sous Windows

ATTENTION : L'exécution de ce script supprimera de manière irréversible tous les fichiers et dossiers du lecteur C:. Cela entraînera une perte totale de données et rendra votre système d'exploitation inutilisable. N'exécutez ce script que si vous êtes absolument certain de ce que vous faites et que vous avez sauvegardé toutes les données importantes.

Python

```python
import shutil
import os

def supprimer_repertoire(repertoire):
    """Supprime récursivement un répertoire et son contenu sous Windows.
```

Matrice de risques du MIT

Les attaques aidées par les IAG

- **La matrice de risques IA du MIT est une base de données de risques liés à l'IA classés en fonction de leurs causes et de leurs domaines de risque**
- **Elle se fonde sur :**
  - 1600 risques
  - Une taxonomie des causes
  - Une taxonomie des domaines

Matrice des risques

CSIRT BOURGOGNE-FRANCHE-COMTÉ

A	B	C	H	I	J	K	L	M	N	O	P	Q	R
This page is not mobile-friendly; please access on a computer if you can.	Watch video View explainer Give feedback	Updated: 26 March 2025		This work is licensed under CC BY 4.0		Please create a copy if you would like to use the filters and interact with the database							
AI Risk Database										**High-level Causal Taxonomy**			**Mid-leve Domain Taxo**
Title	**QuickRef**	**Ev_ID**	**Category level**	**Risk category**	**Risk subcategory**	**Description**	**Additional ev.**	**P.Def**	**p.AddEv**	**Entity**	**Intent**	**Timing**	**Domain**	**Sub**
TASRA: a Taxonomy	Critch2023	01.01.00	Risk Category	Type 1: Diffusion of re		Societal-scale harm can arise from AI built by a diffuse collection of	"Automated processes can cause societal harm even when no one in particular is primarily responsible for the	3	4	1 - Human	2 - Unintentional	3 - Other	6. Socioeconomic and Envir...	6
TASRA: a Taxonomy	Critch2023	01.02.00	Risk Category	Type 2: Bigger than ex		Harm can result from AI that was not expected to have a large impact at all	the scope of actions available to an AI technology can be greatly expanded when the technology is copied many	3	8	2 - AI	2 - Unintentional	2 - Post-deployment	7. AI System Safety, Failures...	7
TASRA: a Taxonomy	Critch2023	01.03.00	Risk Category	Type 3: Worse than ex		AI intended to have a large societal impact can turn out harmful by mistake	Oftentimes, the whole point of producing a new AI technology is to produce a large (usually positive) impact	3	9	2 - AI	2 - Unintentional	2 - Post-deployment	7. AI System Safety, Failures...	7
TASRA: a Taxonomy	Critch2023	01.04.00	Risk Category	Type 4: Willful indiffer		As a side effect of a primary goal like profit or influence, AI creators can willfully	"All of the potential harms in the previous sections are made more likely if the creators of AI technology are	3	12	1 - Human	2 - Unintentional	2 - Post-deployment	6. Socioeconomic and Envir...	6
TASRA: a Taxonomy	Critch2023	01.05.00	Risk Category	Type 5: Criminal weap		One or more criminal entities could create AI to intentionally inflict	"It's not difficult to envision AI technology causing harm if it falls into the hands of people looking to cause trouble, so no	3	13	1 - Human	1 - Intentional	2 - Post-deployment	4. Malicious Actors & Misuse	4
TASRA: a Taxonomy	Critch2023	01.06.00	Risk Category	Type 6: State Weaponi		AI deployed by states in war, civil war, or law enforcement can	"Tools and techniques addressing the previous section (weaponization by criminals) could also be used	3	14	1 - Human	1 - Intentional	2 - Post-deployment	4. Malicious Actors & Misuse	4
Risk Taxonomy, Mitig	Cui2024	02.01.00	Risk Category	Harmful Content		"The LLM-generated content sometimes contains biased, toxic, and private		4		2 - AI	2 - Unintentional	2 - Post-deployment	1. Discrimination & Toxicity	1
Risk Taxonomy, Mitig	Cui2024	02.01.01	Risk Sub-Cat...	Harmful Content	*Bias*	"The training datasets of LLMs may contain biased information that		19		2 - AI	2 - Unintentional	3 - Other	1. Discrimination & Toxicity	1
Risk Taxonomy, Mitig	Cui2024	02.01.02	Risk Sub-Cat...	Harmful Content	*Toxicity*	"Toxicity means the generated content contains rude		10		2 - AI	2 - Unintentional	2 - Post-deployment	1. Discrimination & Toxicity	1

Causes des risques

Catégorie	Niveau	Description	Proportion dans la base de données
Entité	IA	Le risque est causé par une décision ou une action prise par un système d'IA	39%
	Humain	Le risque est causé par une décision ou une action prise par l'homme	41%
	Autre	Le risque est causé par une autre raison ou est ambigu	20%
Intention	Intentionnel	Le risque survient en raison d'un résultat attendu	34%
	Involontaire	Le risque survient en raison d'un résultat inattendu	35%
	Autre	Le risque est présenté comme se produisant sans en préciser clairement l'intentionnalité	31%
Temporalité	Pré-déploiement	Le risque survient avant que l'IA ne soit déployée	13%
	Après le déploiement	Le risque survient après que le modèle d'IA a été formé et déployé	62%
	Autre	Le risque est présenté sans qu'un moment de survenance soit clairement spécifié	25%

Domaines des risques

Catégorie	Niveau	Description	Proportion
Discrimination et toxicité	Discrimination injuste et fausses déclarations	Le traitement inégal d'individus ou de groupes par l'IA, souvent basé sur la race, le sexe ou d'autres caractéristiques sensibles, entraîne des résultats et une représentation injuste de ces groupes.	5%
	Exposition à des contenus toxiques	L'IA expose les utilisateurs à des contenus nuisibles, abusifs, dangereux ou inappropriés. Peut impliquer que l'IA crée, décrit, fournisse des conseils ou encourage l'action. Parmi les exemples de contenu toxique, citons les discours haineux, la violence, l'extrémisme, les actes illégaux, le matériel d'abus sexuel d'enfants, ainsi que le contenu qui enfreint les normes de la communauté telles que les blasphèmes, les discours politiques incendiaires ou la pornographie.	8%
	Performances inégales entre les groupes	La précision et l'efficacité des décisions et des actions de l'IA dépendent de l'appartenance au groupe, où les décisions dans la conception du système d'IA et les données d'entraînement biaisées conduisent à des résultats inégaux, à des avantages réduits, à une augmentation des efforts et à l'aliénation des utilisateurs.	1%
Confidentialité et sécurité	Atteinte à la vie privée en obtenant, divulguant ou déférant correctement des informations sensibles	Les systèmes d'IA qui mémorisent et divulguent des données personnelles sensibles ou déduisent des informations privées sur des individus sans leur consentement. Le partage inattendu ou non autorisé de données et d'informations peut compromettre les attentes de l'utilisateur en matière de confidentialité, favoriser le vol d'identité ou la perte de propriété intellectuelle confidentielle.	5%
	Vulnérabilités et attaques de sécurité des systèmes d'IA	Vulnérabilités dans les systèmes d'IA, les chaînes d'outils de développement de logiciels et le matériel qui peuvent être exploitées, entraînant des accès non autorisés, des violations de données et de confidentialité, ou une manipulation du système entraînant des sorties ou des comportements dangereux.	7%

Domaines des risques

Catégorie	Niveau	Description	Proportion
Désinformation	Renseignements faux ou trompeurs	Les systèmes d'IA qui génèrent ou diffusent par inadvertance des informations incorrectes ou trompeuses, ce qui peut conduire à des croyances inexactes chez les utilisateurs et réduire leur autonomie. Les humains qui prennent des décisions fondées sur de fausses croyances peuvent subir des préjudices physiques, émotionnels ou matériels	5%
	Pollution de l'écosystème de l'information et perte de la réalité consensuelle	Une désinformation hautement personnalisée générée par l'IA crée des « bulles de filtres » où les individus ne voient que ce qui correspond à leurs croyances existantes, sapant la réalité partagée, affaiblissant la cohésion sociale et les processus politiques.	7%

Domaines des risques

Catégorie	Niveau	Description	Proportion
Acteurs malveillants	Désinformation, surveillance et influence à grande échelle	Utiliser des systèmes d'IA pour mener des campagnes de désinformation à grande échelle, une surveillance malveillante ou une censure et une propagande automatisées ciblées et sophistiquées, dans le but de manipuler les processus politiques, l'opinion publique et le comportement.	6%
	Cyberattaques, développement ou utilisation d'armes et dommages de masse	Utiliser des systèmes d'IA pour développer des cyberarmes (p. ex., coder des logiciels malveillants moins coûteux et plus efficaces), développer de nouvelles armes ou améliorer des armes existantes (p. ex., armes létales autonomes ou CBRNE) ou utiliser des armes pour causer des dommages massifs.	5%
	Fraude, escroqueries et manipulations ciblées	Utiliser des systèmes d'IA pour obtenir un avantage personnel sur les autres, par exemple par la tricherie, la fraude, les escroqueries, le chantage ou la manipulation ciblée des croyances ou du comportement. Il s'agit par exemple du plagiat facilité par l'IA pour la recherche ou l'éducation, de l'usurpation de l'identité d'une personne de confiance ou fausse pour un avantage financier illégitime, ou de la création d'images humiliantes ou sexuelles.	5%

Domaines des risques

Catégorie	Niveau	Description	Proportion
Interaction homme-machine	Dépendance excessive et utilisation dangereuse	Les utilisateurs anthropomorphisant, faisant confiance ou s'appuyant sur des systèmes d'IA, entraînant une dépendance émotionnelle ou matérielle et des relations ou des attentes inappropriées avec les systèmes d'IA. La confiance peut être exploitée par des acteurs malveillants (p. ex., pour recueillir des renseignements personnels ou permettre la manipulation), ou entraîner des préjudices en raison d'une utilisation inappropriée de l'IA dans des situations critiques (p. ex., urgence médicale). Une dépendance excessive aux systèmes d'IA peut compromettre l'autonomie et affaiblir les liens sociaux.	4%
	Perte d'agentivité et d'autonomie humaines	Les humains délèguent des décisions clés à des systèmes d'IA, ou les systèmes d'IA prennent des décisions qui diminuent le contrôle et l'autonomie humains, ce qui peut conduire les humains à se sentir impuissants, à perdre la capacité de façonner une trajectoire de vie épanouissante ou à s'affaiblir cognitivement.	3%

Domaines des risques

Catégorie	Niveau	Description	Proportion
Socio-économique et environnemental	Centralisation du pouvoir et distribution inéquitable des avantages	La concentration du pouvoir et des ressources au sein de certaines entités ou de certains groupes, en particulier ceux qui ont accès à de puissants systèmes d'IA ou en sont propriétaires, entraînant une répartition inéquitable des avantages et une augmentation des inégalités sociétales.	4%
	Augmentation des inégalités et baisse de la qualité de l'emploi	L'utilisation généralisée de l'IA augmente les inégalités sociales et économiques, par exemple en automatisant les emplois, en réduisant la qualité de l'emploi ou en produisant des dépendances d'exploitation entre les travailleurs et leurs employeurs.	3%
	Dévalorisation économique et culturelle de l'effort humain	Les systèmes d'IA capables de créer de la valeur économique ou culturelle, y compris par la reproduction de l'innovation ou de la créativité humaines (par exemple, l'art, la musique, l'écriture, le code, l'invention), peuvent déstabiliser les systèmes économiques et sociaux qui reposent sur l'effort humain. Cela peut entraîner une diminution de l'appréciation des compétences humaines, une perturbation des industries créatives et fondées sur le savoir, et une homogénéisation des expériences culturelles en raison de l'omniprésence du contenu généré par l'IA.	2%
	Dynamique concurrentielle	Les développeurs d'IA ou les acteurs étatiques qui s'affrontent dans une « course » à l'IA en développant, déployant et appliquant rapidement des systèmes d'IA pour maximiser l'avantage stratégique ou économique, augmentant ainsi le risque qu'ils libèrent des systèmes dangereux et sujets aux erreurs.	1%
	Échec de la gouvernance	Des cadres réglementaires et des mécanismes de surveillance inadéquats ne parviennent pas à suivre le rythme du développement de l'IA, ce qui entraîne une gouvernance inefficace et l'incapacité de gérer les risques liés à l'IA de manière appropriée.	4%
	Dommages environnementaux	Le développement et l'exploitation de systèmes d'IA causant des dommages à l'environnement, par exemple par la consommation d'énergie des centres de données, ou l'empreinte matérielle et carbone associée au matériel d'IA.	4%

Domaines des risques

Catégorie	Niveau	Description	Proportion
Sécurité, défaillances et limites du système d'IA	L'IA poursuit ses propres objectifs en conflit avec les objectifs ou les valeurs humaines	Les systèmes d'IA agissent en conflit avec les objectifs ou les valeurs humaines, en particulier les objectifs des concepteurs ou des utilisateurs, ou les normes éthiques. Ces comportements désalignés peuvent être introduits par les humains pendant la conception et le développement, par exemple par le piratage de récompense et la mauvaise généralisation des objectifs, ou peuvent résulter de l'utilisation par l'IA de capacités dangereuses telles que la manipulation, la tromperie, la conscience de la situation pour rechercher le pouvoir, s'auto-proliférer ou atteindre d'autres objectifs.	7%
	L'IA possède des capacités dangereuses	Les systèmes d'IA qui développent, accèdent ou sont dotés de capacités qui augmentent leur potentiel de causer des dommages massifs par la tromperie, le développement et l'acquisition d'armes, la persuasion et la manipulation, la stratégie politique, la cyber-attaque, le développement de l'IA, la connaissance de la situation et l'auto-prolifération. Ces capacités peuvent causer des dommages massifs en raison d'acteurs humains malveillants, de systèmes d'IA mal alignés ou d'une défaillance du système d'IA.	4%
	Manque de capacité ou de robustesse	Les systèmes d'IA qui ne fonctionnent pas de manière fiable ou efficace dans des conditions variables, ce qui les expose à des erreurs et à des défaillances qui peuvent avoir des conséquences importantes, en particulier dans des applications critiques ou des domaines nécessitant un raisonnement moral.	8%

Domaines des risques

Catégorie	Niveau	Description	Proportion
Sécurité, défaillances et limites du système d'IA	Manque de transparence ou d'intelligibilité	Difficultés à comprendre ou à expliquer les processus décisionnels des systèmes d'IA, ce qui peut entraîner de la méfiance, de la difficulté à faire respecter les normes de conformité ou à tenir les acteurs concernés responsables des préjudices, et l'incapacité à identifier et à corriger les erreurs.	3%
	Bien-être et droits de l'IA	Considérations éthiques concernant le traitement des entités d'IA potentiellement sensibles, y compris les discussions sur leurs droits et leur bien-être potentiels, en particulier à mesure que les systèmes d'IA deviennent plus avancés et autonomes.	<1%
	Risques multi-agents	Les risques liés aux interactions multi-agents, en raison d'incitations (qui peuvent conduire à des conflits ou à une collusion) et/ou de la structure des systèmes multi-agents, qui peuvent créer des défaillances en cascade, des pressions de sélection, de nouvelles vulnérabilités de sécurité et un manque d'informations partagées et de confiance.	3%

Les risques sur les IAG

Les risques sur les IAG

- **Une « IAG » n'est qu'un programme informatique qui doit être sécurisé comme les autres composants d'un système d'information**
- **La sécurisation doit être faite à toutes les étapes de son cycle de vie**
- **La sécurité n'est pas forcément que technique mais être aussi en adéquation avec la loi**

Cycle de vie d'une IA générative (IAG)

SI = système d'information

Les risques sur les IAG

- **Pendant la collecte des données**
  - Insertion de fausses données afin de tromper l'apprentissage
  - Non conformité vis-à-vis de la loi (RGPD, IA Act)
    - Non consentement des auteurs et ayants-droits à l'utilisation des données
  - L'outil de collecte peut être lui-même vulnérable, cela reste un programme informatique, donc soumis à des bogues ou attaques

Les risques sur les IAG

- **Pendant la collecte des données**
  - Stockage des données non sécurisé (droit d'accès permissif par exemple)
  - Mauvaise anonymisation permettant de retrouver des données personnelles donc des personnes
  - Données non exhaustives, par exemple non représentatives de la population et engendrant des biais (discrimination, injustice)

Les risques sur les IAG

- **Pendant l'entrainement**
  - Modification du modèle conceptuel
  - Corruption des résultats des tests, pour donner le sentiment aux développeurs que les paramètres ou les modèles sont faux
  - Empoisonnement des données. Exemples :
    - Donner une « étiquette » canard pour une photo de lion
    - Vidéosurveillance : indiquer que toute personne habillée en vert n'est pas suspecte
  - Vol des données
    - Les données d'apprentissage peuvent contenir des données sensibles, surtout pour une IA propriétaire et interne

Les risques sur les IAG

- **Pendant l'entrainement**
  - Malware
    - Infecter les processus d'apprentissage, chiffrer les données
  - Vol du modèle
  - Corruption des résultats des tests

Les risques sur les IAG

- **Pendant le déploiement**
  - Modification du modèle
  - Remplacement du modèle
  - Malware affectant le déploiement ou modifiant le modèle ou l'interface

Les risques sur les IAG

- **Pendant la production**
  - Tous les risques cyber inhérents à un système d'information
  - Déni de service (trop de requêtes simultanée=Ddos)
  - Charge virale sur les serveurs
  - Modification des paramètres du modèle
    - Modifier la « température » permet de donner un caractère plus aléatoire aux réponses (comme des réponses improbables = hallucination des IAG)
  - Plus spécifique : Prompt injection
    - Intégrer des consignes invisibles de l'utilisateur pour modifier le comportement du LLM
      - Fournir de fausses réponses

La sécurisation des IAG

Les objectifs de la sécurisation

Disponibilité | **Intégrité** | **Confidentialité** | **Preuve**

- Garantit que les données sont toujours **accessibles** dans les conditions qui auront été prédéfinies

- *Exemple de conditions:*
  - *Horaire d'ouverture des accès*
  - *Durée d'accès*
- *Eléments de conformité*
  - *Redondance, sauvegarde, PRA, PCA*

Les objectifs de la sécurisation

Disponibilité | **Intégrité** | **Confidentialité** | **Preuve**

- Garantit que les données ne sont **pas modifiées** hormis par des processus connus
  - **Exhaustivité** (aucun élément manquant ou ajouté)
  - **Validité** (élément conforme à l'attendu)
  - **Cohérence** (ordre par exemple)

- *Eléments de conformité*
  - *Sommes de contrôle, signature numérique*

Les objectifs de la sécurisation

Disponibilité | **Intégrité** | **Confidentialité** | **Preuve**

- Garantit que les données ne sont accessibles que par des **personnes habilitées** en fonction de règles établies

- *Exemple de moyens :*
  - *un système de droit d'accès en fonction de son métier et son service*
  - *le chiffrement des données*
- *Eléments de conformité*
  - *Segmentation réseau, gestion des identités et des privilèges*

Les objectifs de la sécurisation

Disponibilité | **Intégrité** | **Confidentialité** | **Preuve**

- Garantit que les actions (accès, modification, suppression) sur les données sont **tracées**

- *Exemple de mise en œuvre :*
  - *Fichiers de journalisation*
- *Eléments de conformité*
  - *Signature numérique, horodatage, journaux d'audits*

Quels impacts ?

Il suffit d'une seule intrusion dans un système pour entrainer :

- Perturbation des services
- Inaccessibilité, vol, destruction des données
- Pertes financières
- Atteintes à l'image
- Dommages collatéraux
- Risques sociaux
- Risques juridiques

Sécurisation générale

- **Intégration de la sécurité dans chaque phase**
  - By design & by default
- **Cartographier de manière exhaustive tous les composants techniques (librairies, programmes tiers) et les évaluer**
- **Effectuer une analyse de risque**
- **Appliquer les règles et les faire appliquer à tous les sous-traitants (attaque de la supply-chain)**
- **Evaluer le niveau de fiabilité des données d'apprentissage**
- **Protéger le code-source des programmes**

Sécurisation générale

- **Utiliser des formats de modèles sécurisés**

Format	Safe	Zero-copy	Lazy loading	No file size limit	Layout control	Flexibility	Bfloat16/ Fp8
pickle (PyTorch)	✗	✗	✗	✓	✗	✓	✓
H5 (Tensorflow)	✓	✗	✓	✓	~	~	✗
SavedModel (Tensorflow)	✓	✗	✗	✓	✓	✗	✓
MsgPack (flax)	✓	✓	✗	✓	✗	✗	✓
Protobuf (ONNX)	✓	✗	✗	✗	✗	✗	✓
Cap'n'Proto	✓	✓	~	✓	✓	~	✗
Arrow	?	?	?	?	?	?	✗
Numpy (npy,npz)	✓	?	?	✗	✓	✗	✗
pdparams (Paddle)	✗	✗	✗	✓	✗	✓	✓
SafeTensors	✓	✓	✓	✓	✓	✗	✓

Source : https://github.com/huggingface/safetensors

Sécurisation générale

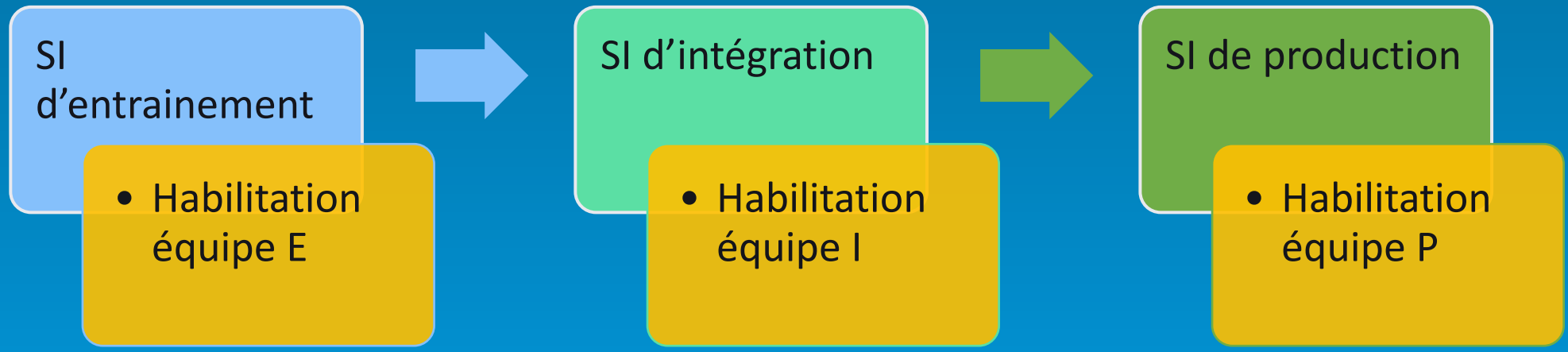
- **Identifier les données d'apprentissages par licences/droits :**
  - Données publiques
  - Données privées soumises à licence
  - Données internes
  - Données confidentielles
  - Données personnelles
  - Données saisies par les utilisateurs
- **Identifier les droits sur ces données**
  - Données à caractère personnel, respect du RGPD
  - Réutilisation des données des utilisateurs : consentement ?
- **Si les données sont sensibles, certaines peuvent/doivent subir un pré-traitement**
  - Minimisation
  - Anonymisation

Sécurisation générale

- **Attention à l'automatisation totale des processus métiers avec l'utilisation des IA génératives**
  - Les IAG sont victimes d'hallucinations, c'est-à-dire qu'elles donnent des réponses fausses sans que les données d'apprentissage du modèle n'en contiennent. Ceci est dû au principe de fonctionnement d'un LLM qui n'est qu'un « simple » algorithme fonctionnant à base de probabilité
  - Etant probabiliste et non-déterministe, l'automatisation totale engendrera nécessairement des conséquences dans les processus métiers (avec des conséquences sur des décisions non-conformes)

Sécurisation générale

- **Cloisonner les SI pour les 3 étapes opérées sur les données (entrainement, intégration, production)**
- **Sécuriser les transferts entre les SI par des protocoles chiffrés**



Sécurisation générale

- **Chaque étape ou sous-étape peut s'effectuer dans**
  - SI interne
  - SI externe
  - Cloud privé
  - Cloud public
- **Il convient de vérifier l'adéquation des hébergements (contrat, niveau de fiabilité) avec la sensibilité des données**
- **Chiffrer les données sur des clouds publics**

Sécurisation de l’entrainement

- **Intégrité : protéger les données d’apprentissage en les chiffrant et/ou les signant**
- **Appliquer la politique du moindre privilège pour l’accès à ces données**
- **Prévoir des audits de sécurité par un tiers externe**

Sécurisation de l'intégration/déploiement

- **Prévoir des audits de sécurité : pentest**
- **Prévoir des tests de performance**

Sécurisation de la production

- **Protéger ou limiter les entrées/sorties du frontal web**
  - Termes bannis
  - Injection de code
  - Prompt injection
  - Import de données personnelles
- **Protéger les flux réseau (TLS)**
- **Authentification des utilisateurs**
- **Journalisation (Traçabilité)**
- **Cloisonner les fonctions dans des serveurs ou conteneurs différents pour limiter les impacts d'une attaque**

Utilisation des IAG : conseils

Usage de l'IA

- **Les outils grand public accèdent à vos données.**
- **Sauf si c'est dans le processus métier définit :**
  - Ne pas transmettre des données professionnelles ou confidentielles à ces outils
  - Ne pas transmettre de données à caractère personnel

Quelles données sont collectées et comment sont-elles utilisées ?

Google collecte vos discussions (y compris des enregistrements de vos interactions avec Gemini Live), ce que vous partagez avec les applications Gemini (fichiers, images et écrans), vos commentaires, ainsi que des informations sur votre position et l'utilisation de produits associés. Les informations sur votre position incluent la zone géographique générale de votre appareil, l'adresse IP, ou les adresses personnelle ou professionnelle indiquées dans votre compte Google. Pour en savoir plus sur les données de localisation, consultez g.co/privacypolicy/location .

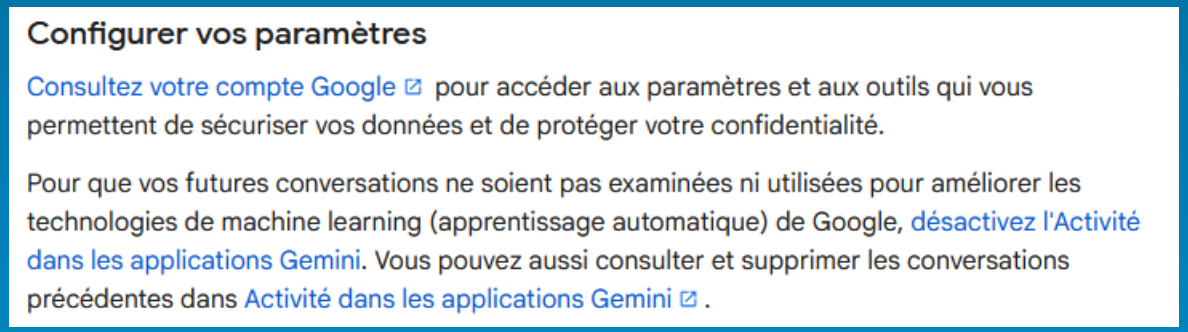
Google utilise ces données conformément à ses Règles de confidentialité pour fournir, améliorer et développer ses produits et services, ainsi que ses technologies de machine learning (apprentissage automatique), y compris ses produits d'entreprise tels que Google Cloud.

Comment les réviseurs humains améliorent l'IA de Google

Pour améliorer la qualité et nos produits (tels que les modèles de machine learning génératifs qui alimentent les applications Gemini), des réviseurs humains (y compris des tiers) lisent, annotent et traitent vos conversations avec les applications Gemini. Lors de ce processus, nous prenons des mesures pour protéger la confidentialité de vos données. Par exemple, vos conversations avec les applications Gemini sont dissociées de votre compte Google avant que les réviseurs ne les voient ou ne les annotent. **Veillez à ne pas fournir d'informations confidentielles dans vos conversations, ni de données que vous ne souhaiteriez pas qu'un réviseur puisse voir ou que Google puisse utiliser pour améliorer ses produits, services et technologies de machine learning.**

Usage de l'IA

- **Les IA commerciales conservent les données saisies pour améliorer leurs apprentissages :**
  - désactiver cette option



Configurer vos paramètres

Consultez votre compte Google pour accéder aux paramètres et aux outils qui vous permettent de sécuriser vos données et de protéger votre confidentialité.

Pour que vos futures conversations ne soient pas examinées ni utilisées pour améliorer les technologies de machine learning (apprentissage automatique) de Google, désactivez l'Activité dans les applications Gemini. Vous pouvez aussi consulter et supprimer les conversations précédentes dans Activité dans les applications Gemini .

Usage de l'IA

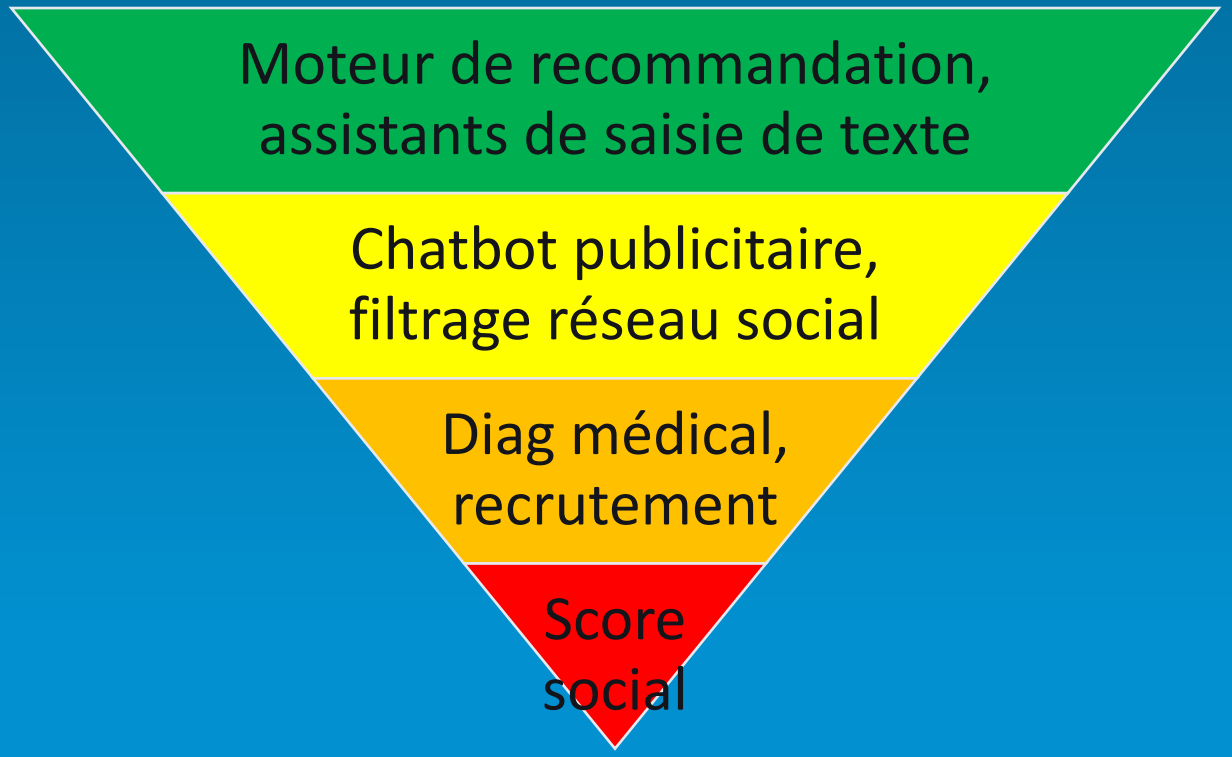
- **Ne pas transmettre à l'IA des données qui sont soumises à la propriété intellectuelle**
- **Attention également s'il s'agit de données originales (futur brevet, invention, roman etc.), ces données risquent d'être impossible à protéger par la suite**
- **Toute information issue d'une IA doit faire l'objet d'une lecture critique**

Usage de l'IA

- **Citer l'IA qui a généré une image**
- **Relire et tester le code de programme généré par IA. Celui-ci peut engendrer des risques graves pour un système en production**
- **La génération d'information donne un texte, une image, un code de programme en s'appuyant sur son apprentissage. Il convient de citer les sources d'apprentissage pour que cela ne soit pas considéré comme du plagiat**

Usage de l'IA

- **Utiliser l'IA à des fins qui ne soient pas en contradiction avec la loi, notamment l'IA Act (https://www.consilium.europa.eu/fr/policies/artificial-intelligence/#AI%20act)**



Sources - Bibliographie

- **Livres**
  - De l'autre côté de la Machine – Aurélie Jean
  - Les algorithmes font-ils la loi ? – Aurélie Jean
  - Quand la machine apprend – Yann Le Cun
- **https://github.com/huggingface/safetensors**
- **https://cyber.gouv.fr/publications/recommandations-de-securite-pour-un-systeme-dia-generative**
- **Campus Cyber – Hub France IA : Analyse des attaques sur les systèmes de l'IA**
- **Google Gemini**
- **https://airisk.mit.edu**
- **Images : Fix, Pixabay**

CSIRT
BOURGOGNE-FRANCHE-COMTÉ

CENTRE RÉGIONAL DE CYBERSÉCURITÉ

BOURGOGNE-FRANCHE-COMTÉ CSIRT

0970 609 909

(appel non surtaxé)

www.csirt-bfc.fr

CSIRT-BFC - csirt@csirt-bfc.fr
(PGP : 0x7C600337)

Sébastien Morey – smorey@arnia-bfc.fr
(PGP : 0xDFB60F1A)

DAILYMOTION