

Intelligence Artificielle et Sécurité Nationale

Lê Nguyễn Hoàng, Calicarpa
Journée des développeurs, DGA MI
Octobre 2024



Section 1

Une nouvelle technologie de l'**information**

Attention à la hype



Personnalisation de services

The image shows a screenshot of the YouTube FR homepage. At the top, there is a search bar with the text "Rechercher" and a magnifying glass icon. To the right of the search bar is a "SE CONNECTER" button. Below the search bar is a navigation bar with various categories: "Tous", "En direct", "Musique", "Chill-out", "Squeezeie", "Jeux vidéo", "Destinations touristiques", "Rires", "Listes de lecture", "Mister V", "Jazz", "Albums", "Comédie à sketches", and "Restauration".

The main content area is titled "Recommended by Tournesol" and features a grid of video recommendations. The first row contains four videos:

- The Future of Public Health: Crash Course Public Health #10** (CrashCourse, 12:50, 44 comparisons by 3 contributors)
- No end to protests over Amini's death** (DW News, 09:24, 48 comparisons by 5 contributors)
- Breaking The Giants** (TechAltar, 15:52, 32 comparisons by 2 contributors)
- Truth Decay** (MinuteEarth, 05:13, 59 comparisons by 18 contributors)

The second row contains four more video recommendations:

- Denzel menace calmement un caïd de la mafia russe | Equalizer | Extrait VF** (Boxoffice, 5:11, 396 k vues, il y a 5 mois)
- Résumé : Avec un grand Benzema, le Real Madrid remporte le Clasico !** (beIN SPORTS France, 10:22, 838 k vues, il y a 20 heures)
- The Good Life Radio • 24/7 Live Radio | Best Relax House, Chillout, Study,...** (The Good Life Radio x Sensual Musique, 10 k spectateurs, EN DIRECT)
- Live Fall Guys PP Venez !! (FACECAM)** (Mr Panditix, 4:34:11, 790 vues, Diffusé il y a 1 jour)

On the left side of the page, there is a vertical navigation menu with icons for "Accueil", "Explorer", "Shorts", "Abonnements", "Bibliothèque", and "Historique".

Détection d'anomalie et de fraude



Généralisation des comportements

 **Translate text**
31 languages

 **Translate files**
.pdf, .docx, .pptx

 **DeepL Write** 
AI-powered edits

French (detected) ▼

 German ▼

Automatic ▼

Glossary

Cher client,
Votre dernière transaction à destination de Nestlé d'un montant de
14 millions de dollars a bien été validée par nos services.
Nous vous remercions pour votre fidélité.
Cordialement.



Sehr geehrter Kunde!
Ihre letzte Transaktion an Nestlé in Höhe von 14 Millionen US-Dollar
wurde von uns bestätigt.
Wir danken Ihnen für Ihre Treue.
Mit freundlichen Grüßen.

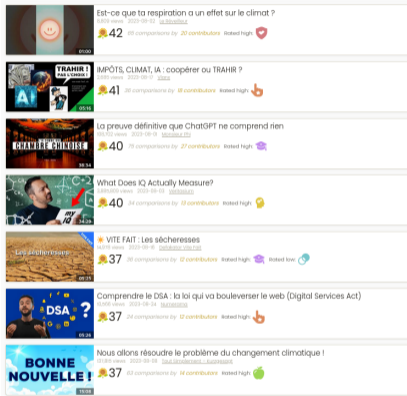


Organiser les documents de votre organisation à l'aide d'une **base de données vectorielles** optimisée pour la **recherche rapide** de plus proche voisins (**RAG**).

Section 2

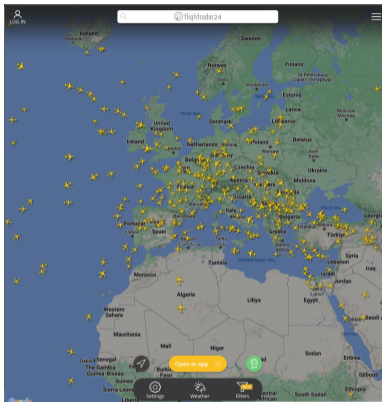
A-t-on vraiment besoin **d'apprentissage** ?

Lequel de ces problèmes est le plus dur ?



A screenshot of a content recommendation feed. It contains six items, each with a thumbnail, title, view count, date, and engagement metrics (comparisons and contributors). The items are:

- Est-ce que ta respiration a un effet sur le climat ?**
1000 views, 2023-08-01, 42 comparisons by 20 contributors, Rated High.
- IMPÔTS, CLIMAT, IA : coopérer ou TRAHIR ?**
995 views, 2023-08-01, 41 comparisons by 48 contributors, Rated High.
- La preuve définitive que ChatGPT ne comprend rien**
16170 views, 2023-08-01, 40 comparisons by 27 contributors, Rated High.
- What Does IQ Actually Measure?**
1400 views, 2023-08-01, 40 comparisons by 49 contributors, Rated High.
- VITE FAIT : Les sécheresses**
1400 views, 2023-08-01, 37 comparisons by 47 contributors, Rated High, Rated Low.
- Comprendre le DSA : la loi qui va bouleverser le web (Digital Services Act)**
11000 views, 2023-08-01, 37 comparisons by 47 contributors, Rated High.
- Nous allons résoudre le problème du changement climatique !**
10180 views, 2023-08-01, 37 comparisons by 46 contributors, Rated High.



Recommandation de contenus VS optimisation du trafic aérien VS échecs

Complexité en temps de calcul

Requiert beaucoup de calculs pour être résolu.

L'optimisation et la cryptanalyse sont de ce type.

Complexité en temps de calcul

Requiert beaucoup de calculs pour être résolu.
L'optimisation et la cryptanalyse sont de ce type.

Complexité en espace de calculs

Requiert beaucoup de "brouillons" pour être résolu.
Les jeux combinatoires (échec, go, poker...) sont de ce type.

Différentes mesure de complexité

Complexité en temps de calcul

Requiert beaucoup de calculs pour être résolu.
L'optimisation et la cryptanalyse sont de ce type.

Complexité en espace de calculs

Requiert beaucoup de "brouillons" pour être résolu.
Les jeux combinatoires (échec, go, poker...) sont de ce type.

Complexité en lignes de code (Kolmogorov-Solomonoff)

Requiert un grand nombre de lignes de code pour être résolu.
Le profilage psychologique, l'analyse d'images et la génération de texte sont de ce type.
Un humain ne peut plus écrire les algorithmes qui résolvent ces problèmes.

L'article révolutionnaire de Turing (1950)

Turing 1950

“C'est surtout un problème de programmation...

Les estimations de la capacité de stockage du cerveau humain varient entre 10^{10} et 10^{15} bits ... Je serais surpris si plus de 10^9 bits étaient nécessaires pour passer le jeu de l'imitation de manière satisfaisante ... Note: La capacité de l'Encyclopaedia Britannica, 11e édition, est 2×10^9 .”

L'article révolutionnaire de Turing (1950)

Turing 1950

"C'est surtout un problème de programmation...

Les estimations de la capacité de stockage du cerveau humain varient entre 10^{10} et 10^{15} bits ...

Je serais surpris si plus de 10^9 bits étaient nécessaires pour passer le jeu de l'imitation de manière satisfaisante ... Note: La capacité de l'Encyclopaedia Britannica, 11e édition, est 2×10^9 ."

En langage moderne

Le jeu de l'imitation a une complexité en lignes de code de l'ordre de 10^9 bits.

L'article révolutionnaire de Turing (1950)

Turing 1950

“C'est surtout un problème de programmation...

Les estimations de la capacité de stockage du cerveau humain varient entre 10^{10} et 10^{15} bits ... Je serais surpris si plus de 10^9 bits étaient nécessaires pour passer le jeu de l'imitation de manière satisfaisante ... Note: La capacité de l'Encyclopaedia Britannica, 11e édition, est 2×10^9 .”

En langage moderne

Le jeu de l'imitation a une complexité en lignes de code de l'ordre de 10^9 bits.

En langage courant

Aucun algorithm avec moins d'un milliard de lignes de code ne peut pas dialoguer comme un humain.

456

A. M. TURING:

Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's? If this were then subjected to an appropriate course of education one would obtain the adult brain. Presumably the child-brain is something like a note-book as one buys it from the stationers. Rather little mechanism, and lots of blank sheets. (Mechanism and writing are from our point of view almost synonymous.) Our hope is that there is so little mechanism in the child-brain that something like it can be easily programmed. The amount of work in the education we can assume, as a first approximation, to be much the same as for the human child.

Section 3

Les systèmes auto-apprenants **apprennent** :
Les conséquences sur leur cybersécurité

Pro > Software & Services

Samsung workers made a major error by using ChatGPT

News By Lewis Maddison published April 04, 2023

Samsung meeting notes and new source code are now in the wild after being leaked in ChatGPT



(Image credit: Valeriya Zankovych / Shutterstock.com)

Samsung workers have unwittingly leaked top secret data whilst using ChatGPT to help them with tasks.

AI February 4, 2023

Microsoft Warns Employees Not to Share Sensitive Data with ChatGPT



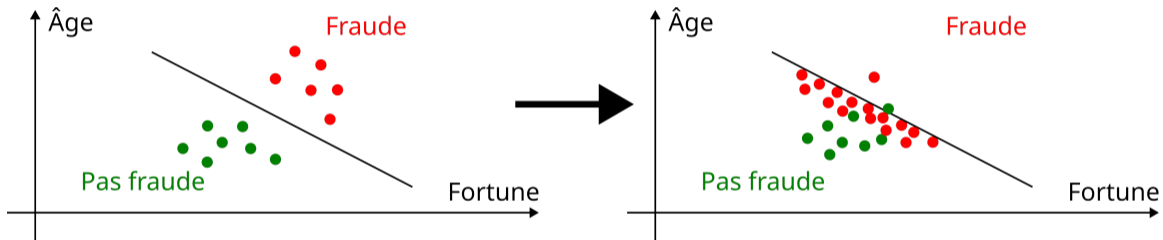
Microsoft has warned its employees not to share sensitive data with an artificially intelligent (AI) chatbot, ChatGPT from OpenAI. Employees of American multinational tech giants had asked in an internal forum whether ChatGPT or any other AI tools from OpenAI were appropriate to use at their work, [Business Insider](#) reported.

Also read: 30% of College Students Use ChatGPT

In response to that inquiry, a senior engineer from Microsoft's CTO office allowed to use ChatGPT but couldn't share confidential information with the AI chatbot.

"Please don't send sensitive data to an OpenAI endpoint, as they may use it for training future models," the senior engineer wrote in an internal post, per Insider.

Évasion (jailbreaking)



Facebook Removed More than 15 Billion Fake Accounts In Two Years, Five Times more than its Active User Base



Jastra Kranjec · Pro Investor 
Updated: 27 September 2021

Disclosure 

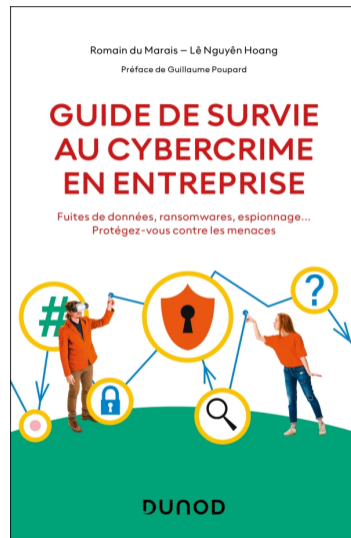
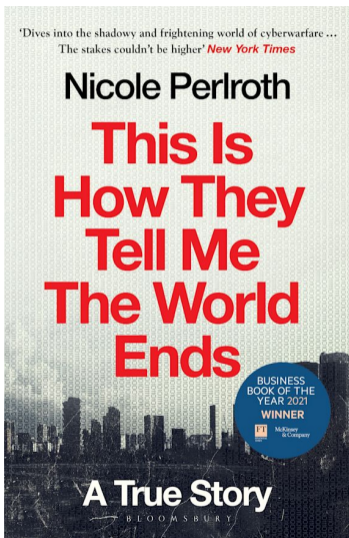
As the world's largest social networking platform, Facebook has witnessed a surge in the number of users in the past few years. Hundreds of millions of people have joined its social media space to communicate, keep in touch with the latest trends or promote business, especially after the pandemic hit. Although the COVID-19 restrictions have loosened in most countries, Facebook's active user base continues growing, but so does the number of fake accounts.

According to data presented by [Stock Apps](#), the social media giant removed over 15 billion fake accounts in the last two years, five times more than its active user base.

3 Billion Fake Accounts Removed in the First Half of 2021, 20x More than the Number of New Active Users

Scammers use fake [Facebook](#) accounts to connect with users, get their personal information and steal identities. Most of them will reach out to anyone who's accepted their friend request to try and scam them out of money.

N'oubliez pas la cybersécurité classique !



Section 4

Stratégies d'atténuation

- Modèles open sources locaux.

- Modèles open sources locaux.
- Sandboxing des algorithmes auto-apprenants.

- Modèles open sources locaux.
- Sandboxing des algorithmes auto-apprenants.
- Confidentialité différentielle.

- Modèles open sources locaux.
- Sandboxing des algorithmes auto-apprenants.
- Confidentialité différentielle.
- Calcul multi-partite.

- Apprentissage adversarial.

- Apprentissage adversarial.
- Détection de données hors-distribution.

- Apprentissage adversarial.
- Détection de données hors-distribution.
- Analyse de glissement distributionnel.

- Apprentissage adversarial.
- Détection de données hors-distribution.
- Analyse de glissement distributionnel.
- Recours pour les faux négatifs/positifs.

- Nettoyage des données.

- Nettoyage des données.
- Authentification (cryptographique) des sources.

- Nettoyage des données.
- Authentification (cryptographique) des sources.
- Apprentissage par agrégations résilientes.

- Nettoyage des données.
- Authentification (cryptographique) des sources.
- Apprentissage par agrégations résilientes.
- Réduction de la dimensionalité.

- Réduire la surface d'attaque.

- Réduire la surface d'attaque.
- Cloisonnement des composants avec moindre privilège.

- Réduire la surface d'attaque.
- Cloisonnement des composants avec moindre privilège.
- Redondance et diversification des systèmes critiques.

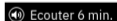
- Réduire la surface d'attaque.
- Cloisonnement des composants avec moindre privilège.
- Redondance et diversification des systèmes critiques.
- Monitoring du système d'information.

Section 5

Le réflexe cyberdéfense :
Qu'en feront les cybercriminels ?

« L'arnaque au président » dopée par un deepfake : cette entreprise s'est fait voler 26 millions de dollars

Par Challenges.fr le 04.02.2024 à 21h17



Une entreprise basée à Hong Kong s'est fait dérober 26 millions de dollars. Les escrocs ont utilisé les mêmes procédés que l'arnaque au président, une escroquerie qui avait touché des milliers de sociétés dans le monde au milieu des années 2010. Mais en 2024, l'IA et les *deepfake* boostent cette pratique frauduleuse.

CAC 40 -0,44% 7467,60

La journaliste Salomé Saqué appelle à la régulation des deepfakes pornographiques



next

Le 12 décembre 2023 à 06h53

IA et algorithmes 2 min

Journaliste pour Blast!, autrice de *Sois jeune et tais-toi* : Réponse à ceux qui critiquent la jeunesse, Salomé Saqué a été visée pour la seconde fois par du deepfake pornographique.

Elle a pris le parti d'en parler publiquement, sur ses réseaux sociaux et auprès du média Fraiches, tant pour sensibiliser au problème que pour appeler à une régulation du phénomène.

En Continu

La gendarmerie aurait mis « plus d'un an » à craquer le cryptophone Ghost

Séou

Le prix de YouTube Premium s'envole dans certains pays

Web

Clara Chappaz, de la French Tech au secrétariat d'État à l'IA et au numérique

Société 9

IA : un impact environnemental conséquent mais toujours difficile à mesurer

IA 2

Qualcomm aurait approché Intel pour un rachat

Éco

Google crée un fonds de 120 millions de dollars pour l'IA dans l'éducation

IA

668e édition des #LIDD : Liens Intelligents Du Dimanche

Next 5

“TEAM JORGE”: IN THE HEART OF A GLOBAL DISINFORMATION MACHINE

In Part 2 of the “Story Killers” project, which continues the work of assassinated Indian journalist Gauri Lankesh on disinformation, the Forbidden Stories consortium investigated an ultra-secret Israeli company involved in manipulating elections and hacking African politicians. **We took an unprecedented dive into a world where troll armies, cyber espionage and influencers are intertwined.**



Les grands modèles de langage : quels risques ? Échange avec Lê Nguyễn Hoang

Paroles de | 23 juin 2023

Chercheur et vulgarisateur scientifique, Lê Nguyễn Hoang revient dans cet entretien sur les enjeux de sécurité, de régulation ou encore de désinformation que pose l'émergence auprès du grand public des intelligences artificielles génératives et de leurs multiples applications.



Aujourd'hui, dans le numérique, nous nous posons la question de la conformité des systèmes avec la loi après leur déploiement. Je défends le principe de présomption de non-conformité.

Pour en revenir aux intelligences artificielles génératives, la taille des modèles de langage me semble être une clé d'entrée intéressante pour le régulateur. Pour faire fonctionner ces modèles, il faut des quantités massives de données et celles-ci sont souvent générées par les utilisateurs eux-mêmes. **En allant vers des modèles plus réduits, qui fonctionnent avec moins de données, nous pourrions avoir des systèmes plus conformes à la loi et plus sécurisés.** Le nombre d'utilisateurs pourrait également être un critère d'attention. Nous pourrions revenir à la présomption de non-conformité dès qu'une application atteint un ordre de grandeur particulier. TikTok en constitue une illustration : l'application devrait être beaucoup plus régulée du fait de l'augmentation de son nombre d'utilisateurs, et donc de son influence [ndlr : TikTok compte aujourd'hui plus d'1,5 milliards d'utilisateurs au total [🔗](#)].

Des armes déjà bien infiltrées en pays rivaux



53

410
comparisons

137
contributors



Éthique 20



30:18

Science4All

TikTok : la machine de propagande la pl...

Section 6

Un risque existentiel ?



Review of the Summer 2023 Microsoft Exchange Online Intrusion

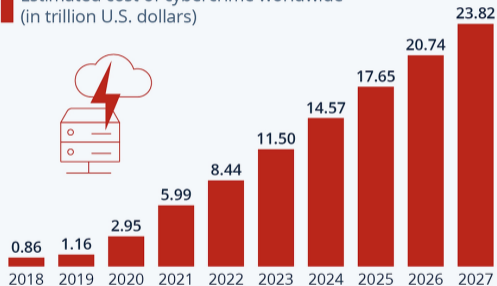
March 20, 2024
Cyber Safety Review Board

The Board is also concerned with Microsoft's public communications after the incident. In its September 6, 2023 blog post entitled "*Results of Major Technical Investigations for Storm-0558 Key Acquisition*," Microsoft explained that Storm-0558 likely stole the 2016 MSA key in the "crash dump" scenario described above. However, soon after publishing that blog, Microsoft determined it did not have any evidence showing that the crash dump contained the 2016 MSA key. This led Microsoft to assess that the crash dump theory was no longer any more probable than other theories as the mechanism by which the actor had acquired the key, which Microsoft chose to leave uncorrected for more than six months after publishing its September 6 blog.

The Board is troubled that Microsoft neglected to publicly correct this known error for many months. Customers (private sector and government) relied on these public representations in Microsoft's blogs. The loss of a signing key is a serious problem, but the loss of a signing key through unknown means is far more significant because it means that **the victim company does not know how its systems were infiltrated and whether the relevant vulnerabilities have been closed off**. Left with the mistaken impression that Microsoft has conclusively identified the root cause of this incident, Microsoft's customers did not have essential facts needed to make their own risk assessments about the security of Microsoft cloud environments in the wake of this intrusion. Microsoft told the Board early in this review that it believed that the errors in the blog were "not material." The Board disagrees. After several written follow up questions from the Board regarding the blog, Microsoft informed the Board on March 5, 2024, that it would be updating the blog in the "near future." One week following this communication, and more than six months after its publication of the September 6 blog, Microsoft corrected its mistaken assertions through an addendum to the blog's existing webpage.

Cybercrime Expected To Skyrocket in the Coming Years

Estimated cost of cybercrime worldwide
(in trillion U.S. dollars)



As of November 2022. Data shown is using current exchange rates.

Sources: Statista Technology Market Outlook,
National Cyber Security Organizations, FBI, IMF



statista

Microsoft is a national security threat, says ex-White House cyber policy director

With little competition at the government level, Windows giant has no incentive to make its systems safer

 [Brandon Vigliarolo](#)

Sun 21 Apr 2024 // 15:25 UTC

INTERVIEW Microsoft has a shocking level of control over IT within the US federal government – so much so that former senior White House cyber policy director AJ Grotto thinks it's fair to call Redmond's recent security failures a national security issue.

Grotto this week spoke with *The Register* in an interview you can watch below, in which he told us that exacting even slight concessions from Microsoft has been a major fight for the Feds.

Previous Reports



DR 2022:
Autocratization
Changing Nature?



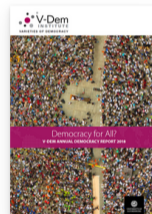
DR 2021:
Autocratization Turns
Viral



DR 2020:
Autocratization Surges -
Resistance Grows



DR 2019: Democracy
Facing Global
Challenges



DR 2018: Democracy
for All?



DR 2017: Democracy at
Dusk?

Les IA de recommandation : plus lucratives et dangereuses que ChatGPT



Pixabay image by LolaSandoval1.

PUBLICATIONS

NIST AI 100-2 E2023

Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations



Date Published: January 2024

Author(s)

Apostol Vassilev (NIST), Alina Oprea (Northeastern University), Alie Fordyce (Robust Intelligence), Hyrum Anderson (Robust Intelligence)

Abstract

This NIST Trustworthy and Responsible AI report develops a taxonomy of concepts and defines terminology in the field of adversarial machine learning (AML). The taxonomy is built on surveying the AML literature and is arranged in a conceptual hierarchy that includes key types of ML methods and lifecycle stages of attack, attacker goals and objectives, and attacker capabilities and knowledge of the learning process. The report also provides corresponding methods for mitigating and managing the consequences of attacks and points out relevant open challenges to take into account in the lifecycle of AI systems. The terminology used in the report is consistent with the literature on AML and is complemented by a glossary that defines key terms associated with the security of AI systems and is intended to assist non-expert readers. Taken together, the taxonomy and terminology are meant to inform other standards and future practice guides for assessing and managing the security of AI systems, by establishing a common language and understanding of the rapidly developing AML landscape.

Keywords

artificial intelligence; machine learning; attack taxonomy; evasion; data poisoning; privacy breach; attack mitigation; data modality; chatbot; generative models; large language model; trojan attack; backdoor attack

Control Families

DOCUMENTATION

Publication:

<https://doi.org/10.6028/NIST.AI.100-2e2023>

[Download URL](#)

Supplemental Material:

[Trustworthy & Responsible AI Resource Center](#)

[NIST news article](#)

Document History:

10/30/19: [IR 8269 \(Draft\)](#)

03/08/23: [AI 100-2 E2023 \(Draft\)](#)

01/04/24: [AI 100-2 E2023 \(Final\)](#)

TOPICS

Security and Privacy

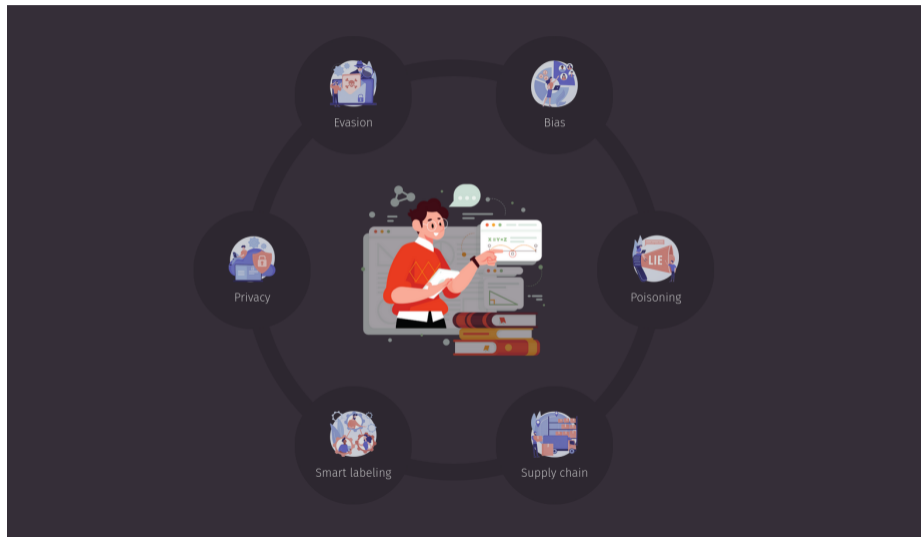
[advanced persistent threats, botnets, information sharing, intrusion detection & prevention, malware](#)

Technologies

[artificial intelligence](#)

5 des 9 articles cités sur l'empoisonnement sont co-écrits par les co-fondateurs de Calicarpa.

Faites de vos collaborateurs des acteurs de la cybersécurité



Section 7

La cybersécurité : en enjeu sociétal

- Exfiltration des données.
- Vulnérabilités des modèles **statistiques** entraînés.
- Empoisonnement de l'apprentissage.
- Utilisations malveillantes par le cybercrime.



The screenshot shows the Calicarpa website with a dark purple background. The navigation bar includes the Calicarpa logo, the name 'Calicarpa', and links for 'Research', 'Consulting', 'Product', 'About Us', and a 'Contact Us' button. The main content area features a large heading 'AI Security' and a sub-heading 'Protect your business against poisoning attacks'. Below this is a text box with three paragraphs of text and an illustration of a person at a workstation with robotic arms and data screens.

Calicarpa

Research Consulting Product About Us Contact Us

AI Security

Protect your business against poisoning attacks

Artificial intelligence creates fantastic opportunities that can drastically increase the added value of your business.

But rushed integrations of **poorly secured** AI systems will expose your business to much greater risks.

Cyber risks must not be underestimated. Recall that the cost of **cybercrime** is evaluated at \$11 trillion in 2023 alone.



L'ambition profonde : rendre le numérique démocratique

