Vincent GENOT

L'Intelligence Artificielle et cybersécurité

Chapitre 2 : Le déploiement

Une dichotomie complexe entre avancées protectrices et vecteurs de menaces émergents

Promotion 2024-2025

Nous avons développé dans le premier chapitre, une approche innovante de la cyberdéfense par l'intégration de la « deceptive » technologie, toutefois, devant l'afflux de données à traiter, et le périmètre numérique et informationnel, nous avons souhaité mettre en place une IA pour nous aider à prédire les nouvelles cybermenaces. Mais avant toute implémentation technique d'une nouvelle technologie comme l'IA, je me devais de m'intéresser également aux nouvelles menaces implicitement émergentes avec la démocratisation de l'IA dans tous nos actifs numériques.

Résumé : L'intelligence artificielle (IA), concept autrefois confiné à la spéculation futuriste, s'est imposée comme une technologie transformatrice aux implications profondes pour la cybersécurité. Cet article examine la dialectique entre les représentations mythiques de l'IA et sa concrétisation actuelle, marquée par une démocratisation rapide via son intégration dans les ordinateurs personnels et les appareils mobiles. Nous analyserons ensuite le rôle ambivalent de l'IA en cybersécurité : d'une part, un puissant levier pour l'optimisation des défenses et la proactivité face aux menaces ; d'autre part, une source de vulnérabilités inédites, touchant des architectures fondamentales, jusqu'aux modèles les plus complexes, devenant un outil potentiellement exploitable à des fins malveillantes. L'objectif est de fournir une perspective éclairée sur les défis et opportunités que l'IA présente pour la sécurité du numérique.

1. L'Intelligence Artificielle : Entre Mythe Persistant, Réalité Opérationnelle et Incarnation Physique Balbutiante

Le concept d'intelligence artificielle (IA) est historiquement imprégné d'une aura mythique, largement façonnée par la littérature et le cinéma de science-fiction. Le **mythe** de l'IA évoque souvent des entités dotées de conscience (une "IA forte" ou Intelligence Artificielle Générale - AGI), d'une autonomie décisionnelle absolue et d'une capacité à surpasser l'intellect humain, *voire à développer une volonté propre pouvant entrer en conflit avec les desseins de ses créateurs*.

- Les "lois de la robotique" d'Isaac Asimov dans son recueil *I, Robot* (1950) [1] tentaient de poser un cadre éthique à des intelligences positroniques complexes, soulignant déjà la tension entre création et contrôle.
- HAL 9000, l'ordinateur de bord de *2001 : L'Odyssée de l'espace* (Kubrick & Clarke, 1968) [2], incarne la déviance d'une IA dont la logique interne la pousse à des actes extrêmes.
- Plus tard, des œuvres comme *Blade Runner* (Scott, 1982, basé sur Philip K. Dick) explorent la frontière floue entre humanité et artificialité, tandis que le genre cyberpunk, avec des pionniers comme William Gibson (*Neuromancer*, 1984) [3], dépeint des IA évoluant en symbiose ou en antagonisme avec les humains au sein de vastes réseaux numériques, le "cyberspace".

Ces narratifs, bien que stimulants, ont contribué à une perception de l'IA souvent éloignée de son état actuel. La figure de **Skynet** dans la franchise *Terminator* (Cameron, 1984) [4], une IA militaire devenant consciente et déclenchant une guerre contre l'humanité, reste un archétype puissant de cette peur de l'IA omnipotente.

Quant à « La vision dystopique » de *1984* de George Orwell (1949) [34], bien que non centrée sur l'IA, préfigure une société de surveillance omniprésente que les technologies actuelles augmentée à l'IA pourraient exacerber.

La **réalité** contemporaine de l'IA, qualifiée d'IA "faible" ou d'Intelligence Artificielle Étroite (ANI), est intrinsèquement différente. Elle repose sur des algorithmes, principalement issus de l'apprentissage automatique (Machine Learning - ML) et de l'apprentissage profond (Deep Learning - DL), qui sont des sous-domaines des statistiques et de l'optimisation mathématique.

Ces systèmes sont conçus pour exécuter des tâches spécifiques avec une efficacité souvent surhumaine, après avoir été entraînés sur de vastes corpus de données. Le ML permet aux machines d'apprendre à partir de données sans être explicitement programmées pour chaque cas, tandis que le DL, utilisant des réseaux de neurones artificiels à multiples couches (dont les Perceptrons Multi-Couches - MCP - sont un exemple fondamental), excelle dans la reconnaissance de motifs complexes (images, son, langage naturel) [5, Bengio Y., Goodfellow I., Courville A., 2016, "Deep Learning"].

Ces IA sont des outils sophistiqués, mais dépourvus de conscience, d'intentionnalité intrinsèque ou d'une compréhension du monde qui transcenderait les données sur lesquelles elles ont été entraînées. Elles excellent dans des domaines pour lesquels elles ont été formées, mais manquent de la polyvalence et de la capacité d'adaptation de l'intelligence humaine générale.

Fondamentalement, l'IA logicielle, telle que nous la développons et l'utilisons massivement aujourd'hui, est **cantonnée au monde virtuel**. Son existence et son fonctionnement dépendent intrinsèquement des **centres de calcul (datacenters)**, des infrastructures informatiques

distribuées (cloud computing) et des réseaux de communication qui lui fournissent la puissance de calcul nécessaire et les données dont elle se nourrit.

Sans ces « **substrats** » physiques et numériques, les grands modèles de langage ou les systèmes de reconnaissance d'image les plus avancés seraient inertes. En ce sens, ***l'IA n'a pas d'existence autonome en dehors de cet écosystème technologique***.

Son **extension dans le monde physique**, c'est-à-dire sa capacité à interagir directement et de manière significative avec notre environnement tangible, reste, malgré des progrès notables, **limitée par notre capacité actuelle à l'intégrer efficacement dans des composants, des dispositifs et des systèmes de la vie de tous les jours**. En effet, cette intégration prend souvent la forme de **"petites IA" ou d'IA embarquées, dédiées à des tâches très spécifiques**. On les retrouve dans les systèmes d'aide à la conduite automobile (ADAS), les assistants vocaux domestiques, les algorithmes optimisant la consommation d'énergie des appareils électroménagers, ou encore les capteurs intelligents dans l'agriculture de précision. Ces IA appelés « edge AI » fonctionnent souvent avec des modèles allégés et optimisés pour des ressources matérielles contraintes.

Cependant, il est crucial de noter que cette **extension de l'IA dans le monde physique est en pleine démocratisation**, un phénomène qui sera détaillé plus loin (voir section 2.2, "L'IA dans nos poches"). L'avènement de puces spécialisées (NPUs, TPUs embarqués) de plus en plus puissantes et économes en énergie permet d'exécuter des inférences IA directement sur les appareils, ouvrant la voie à des applications plus réactives, plus respectueuses de la vie privée (traitement local) et moins dépendantes d'une connexion permanente au cloud. Ainsi, bien que l'IA "générale" fantasmée reste une construction virtuelle, des formes d'intelligence artificielle spécialisée commencent à s'incarner physiquement et à se diffuser dans notre environnement quotidien, non pas comme des entités autonomes, mais comme des fonctionnalités intelligentes intégrées à des objets existants.

En conclusion de cette première partie, et en s'appuyant sur l'analyse de ses architectures, de ses modes opératoires et de ses dépendances infrastructurelles, **l'IA, telle que nous la connaissons et la développons aujourd'hui, opère principalement dans le monde virtuel**, s'appuyant sur des infrastructures numériques massives. Son incarnation physique est progressive, fragmentée en applications spécialisées, et sa démocratisation dans les objets du quotidien est une tendance forte mais encore loin de l'autonomie et de la généralité des IA de la science-fiction. Elle n'a pas d'incarnation physique autonome ni de compréhension du monde réel au-delà des représentations numériques qui lui sont fournies et des tâches pour lesquelles elle a été spécifiquement entraînée et déployée.

2. L'État Actuel de l'IA : Une Vague de Démocratisation et d'Innovation Accélérée

L'écosystème de l'intelligence artificielle a connu une transformation radicale ces dernières années, caractérisée par une démocratisation et une accessibilité sans précédent, déplaçant l'IA des laboratoires de recherche spécialisés vers **les mains du grand public**.

Un jalon emblématique de cette nouvelle ère est la mise à disposition publique de **ChatGPT par OpenAI le 30 novembre 2022** [6, OpenAI Blog, "Introducing ChatGPT"]. Ce grand modèle de langage (LLM), basé sur l'architecture Transformer [7, Vaswani et al., 2017, "Attention Is All You Need"], a non seulement démontré des capacités conversationnelles et de génération de texte

stupéfiantes, mais a surtout sensibilisé le grand public et les entreprises à l'énorme potentiel de l'IA générative, déclenchant une vague d'intérêt et d'investissement à l'échelle mondiale.

Parallèlement à ces avancées propriétaires, l'essor de plateformes collaboratives comme **Hugging Face** a été déterminant [8, Hugging Face, "About Us"]. En centralisant des milliers de modèles pré-entraînés (BERT, GPT, Stable Diffusion, etc.), des jeux de données, et des bibliothèques logicielles (comme transformers et diffusers), a considérablement abaissé la barrière à l'entrée pour les chercheurs, développeurs et entreprises souhaitant expérimenter et intégrer l'IA. Cette dynamique a été amplifiée par la publication de nombreux **modèles open source** puissants, tels que LLaMA et Llama 2 de Meta [9, Touvron et al., 2023, "Llama 2: Open Foundation and Fine-Tuned Chat Models"], BLOOM par le consortium BigScience [10, BigScience Workshop, 2022, "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model"], ou encore les modèles de Mistral AI. Ces initiatives, bien que parfois accompagnées de licences d'utilisation spécifiques, ont stimulées l'innovation et permis une plus grande transparence et auditabilité des modèles, favorisant une adoption plus large.

Cette **vulgarisation** s'est traduite par une prolifération d'**outils librement accessibles ou à bas coût**, incluant des interfaces de programmation (API), des plateformes "low-code/no-code", et des logiciels facilitant l'installation locale et le "fine-tuning" de modèles sur des matériels plus modestes. Un phénomène particulièrement intéressant est l'émergence de **nouveaux services IA développés par d'autres IA**. L'IA est désormais capable de générer du code informatique fonctionnel (ex: GitHub Copilot, AlphaCode [11, Li et al., 2022, "Competition-Level Code Generation with AlphaCode"]), d'optimiser les paramètres d'autres modèles d'IA (AutoML), ou de concevoir des architectures de réseaux de neurones (Neural Architecture Search - NAS).

2.1. L'IA Intégrée au Cœur des Systèmes d'Exploitation Personnels : L'Exemple de Microsoft Copilot et Windows 11

La démocratisation de l'IA franchit une nouvelle étape avec son intégration native au sein des systèmes d'exploitation des ordinateurs personnels. **Microsoft Copilot**, annoncé et progressivement déployé dans **Windows 11** et la suite Microsoft 365, illustre parfaitement cette tendance [36, Microsoft Blog/Announcements on Copilot in Windows 11]. Copilot se positionne comme un assistant IA omniprésent, capable d'interagir avec l'utilisateur en langage naturel pour effectuer une multitude de tâches : résumer des documents, rédiger des emails, générer des images, modifier des paramètres système, ou encore automatiser des flux de travail au sein des applications Office (Word, Excel, PowerPoint, Teams). Cette intégration profonde vise à rendre l'IA accessible de manière contextuelle et intuitive, sans que l'utilisateur ait besoin de naviguer vers une application tierce ou de posséder des compétences techniques spécifiques en IA.

Cette évolution est soutenue par des avancées matérielles, notamment l'émergence de **Processeurs Neuronaux (NPUs - Neural Processing Units)** intégrés directement dans les nouvelles générations de processeurs pour PC (par exemple, chez Intel, AMD, et Qualcomm) [37, Articles techniques sur les NPUs dans les processeurs x86/ARM pour PC]. Ces NPUs sont conçus pour accélérer les calculs liés à l'IA directement sur l'appareil (on-device), offrant plusieurs avantages : performance accrue, meilleure confidentialité par traitement local des données, réactivité améliorée et moindre dépendance à une connexion internet constante. L'intégration de Copilot dans Windows 11, combinée à la disponibilité de matériel optimisé, signifie que des millions d'utilisateurs de PC auront un accès direct et simplifié à des capacités d'IA avancées, transformant potentiellement la manière dont ils interagissent avec leur machine et accomplissent leurs tâches quotidiennes.

2.2. L'IA dans nos Poches : La Révolution Silencieuse des Appareils Mobiles et l'Incarnation Physique de l'IA Spécialisée

Parallèlement aux PC, les appareils mobiles, en particulier les smartphones, sont devenus des vecteurs majeurs de démocratisation de l'IA et une illustration concrète de son **extension dans le monde physique sous forme d'IA dédiées**. Qu'il s'agisse des systèmes **Android de Google** ou **iOS d'Apple**, l'IA est de plus en plus intégrée, souvent de manière transparente pour l'utilisateur, transformant ces appareils en véritables plateformes d'IA personnelles.

Chez **Apple**, l'IA est une composante clé, avec un accent marqué sur le traitement **on-device** grâce à son **Neural Engine**, une partie dédiée de ses puces de la série A et M [38, Documentation Apple sur le Neural Engine]. Cette approche privilégie la performance et la confidentialité. Les manifestations de ces "petites IA" spécialisées sur iOS incluent :

- **Siri :** L'assistant vocal, dont les capacités de compréhension et de réponse en langage naturel sont continuellement améliorées par des modèles de ML exécutés localement pour de nombreuses requêtes.
- **Photographie augmentée par l'IA :** Des fonctionnalités comme le mode Portrait (qui simule un effet bokeh en comprenant la profondeur de la scène), Smart HDR (qui fusionne plusieurs expositions), Deep Fusion (qui analyse les images pixel par pixel), et la reconnaissance de scènes/objets dans l'application Photos. Ce sont des tâches complexes d'IA visuelle qui améliorent drastiquement la qualité des images prises par un appareil compact.
- **Texte en direct (Live Text) et Recherche Visuelle (Visual Look Up) :** Des IA capables de reconnaître et d'extraire du texte ou d'identifier des objets, des plantes, des animaux directement à partir des images capturées par l'appareil photo ou stockées sur l'appareil.
- **Suggestions Intelligentes :** Le clavier prédictif qui anticipe les mots, les suggestions de widgets, d'applications ou d'actions basées sur l'analyse des habitudes de l'utilisateur, le tout traité localement pour préserver la vie privée (ou pas).
- **Fonctionnalités d'Accessibilité :** La reconnaissance sonore pour alerter les personnes malentendantes de sons spécifiques (alarme incendie, sonnette), la description d'images pour les malvoyants, sont des applications directes d'IA spécialisée améliorant l'autonomie.

Du côté d'**Android**, **Google** a également massivement investi dans l'IA, tirant parti de son expertise en la matière pour intégrer des IA dédiées dans l'expérience utilisateur. L'**Assistant Google** est un exemple phare, mais l'IA infuse de nombreuses autres fonctionnalités [39, Blog Google AI ou Android Developers sur les fonctionnalités IA]:

- **Google Lens :** Une IA visuelle puissante capable de reconnaître des objets, du texte, des lieux, des produits via l'appareil photo, offrant des informations contextuelles ou des actions.
- **Smart Reply et Smart Compose :** Des suggestions de réponses rapides et une aide à la rédaction dans Gmail et Messages, générées par des modèles de langage optimisés pour une exécution mobile.
- **Photographie Avancée :** Notamment sur les téléphones Pixel avec des fonctionnalités comme Night Sight (pour les photos en basse lumière), Portrait Mode, et Magic Eraser

(pour supprimer des éléments indésirables des photos), souvent propulsées par les puces **Google Tensor** qui intègrent des cœurs TPU (Tensor Processing Units) dédiés à l'accélération des tâches IA [40, Documentation Google sur les puces Tensor].

- **Live Caption et Enregistreur avec Transcription :** Transcription en temps réel de l'audio (appels, vidéos, podcasts) directement sur l'appareil, rendant le contenu plus accessible.
- **Now Playing (Pixel) :** Reconnaissance automatique de la musique jouée à proximité, fonctionnant entièrement en local grâce à une base de données musicale embarquée et des algorithmes d'IA.
- **Optimisation de la Batterie et des Performances (Adaptive Battery, Adaptive Brightness) :** L'IA est utilisée pour apprendre les habitudes de l'utilisateur et gérer les ressources du système (applications en arrière-plan, luminosité de l'écran) de manière plus intelligente et économe en énergie.

L'intégration de ces "petites IA" spécialisées dans les smartphones, souvent aidée par des co-processeurs IA dédiés, signifie que des milliards d'utilisateurs bénéficient quotidiennement de fonctionnalités améliorées par l'IA sans nécessairement s'en rendre compte. Cela va de l'amélioration de la qualité des photos à des interactions plus intuitives et une meilleure gestion des ressources de l'appareil. Cette **extension de l'IA dans le monde physique via les objets du quotidien** est une tendance forte, rendue possible par la miniaturisation, l'efficacité énergétique croissante des puces IA et la sophistication des modèles capables de fonctionner avec des ressources limitées.

En conclusion de cette section, la convergence de la disponibilité de modèles puissants (souvent open source), de plateformes collaboratives, d'outils de développement simplifiés, et surtout de l'intégration native de l'IA au sein des systèmes d'exploitation des PC et des appareils mobiles (incarnant des IA spécialisées), soutenue par des avancées matérielles significatives, a créé une **opportunité sans précédent pour une vaste frange de la société** d'exploiter la puissance de l'IA. L'intelligence artificielle n'est plus un domaine réservé aux experts, mais devient une composante de plus en plus accessible et intégrée de notre quotidien numérique et physique, ouvrant des perspectives d'innovation disruptive dans tous les secteurs.

3. L'IA en Cybersécurité : Entre Renforcement Défensif et Escalade des Menaces

L'intégration de l'intelligence artificielle dans le champ de la cybersécurité est une lame à double tranchant. D'un côté, elle offre des capacités révolutionnaires pour anticiper, détecter et contrer les cybermenaces. De l'autre, elle introduit de nouvelles surfaces d'attaque et peut être détournée pour créer des menaces plus sophistiquées et évasives.

3.1. L'IA comme Multiplicateur de Force pour la Cyberdéfense

L'IA s'est imposée comme un allié précieux pour augmenter l'efficacité et la réactivité des stratégies de cybersécurité.

- **Détection Améliorée des Menaces :** L'IA, notamment via l'apprentissage non supervisé, excelle dans l'**analyse comportementale des utilisateurs et des entités (UEBA)**. En établissant des lignes de base du comportement normal, elle peut identifier des déviations subtiles indicatives d'une compromission [12, Sans Institute]. L'IA permet aussi une **lecture et reconnaissance automatique** d'objets, de textes et de contextes dans les flux de données massifs.

- **Automatisation et Accélération de la Réponse :** L'IA permet de trier, corréler et prioriser les alertes de sécurité et d'automatiser certaines étapes de la réponse à incident (SOAR), **accélérant la résolution des problèmes** [13, IBM Security].
- **Analyse Prédictive et Découverte de Corrélations :** L'IA peut traiter des volumes massifs de données hétérogènes pour **trouver des nouvelles corrélations** et identifier des schémas d'attaque émergents.
- **Solutions Spécialisées et Réponse Autonome Rapide :** Des entreprises comme **Darktrace** [14] ou **Vectra AI** [15] illustrent l'application concrète de l'IA pour une détection affinée et une réponse rapide.

3.2. Les Dangers Potentiels et l'Instrumentalisation Malveillante de l'IA

Malgré ses contributions, l'omniprésence de l'IA introduit un éventail complexe de nouveaux vecteurs de menace, touchant des architectures comme les MCP jusqu'aux modèles les plus complexes.

- **Surveillance, Contrôle et Érosion de la Vie Privée :**
  La capacité de l'IA à analyser massivement des données personnelles peut mener à une **privation de libertés par plus de contrôles** [25, Foucault]. L'argument *"qui n'a rien à cacher, n'a pas de secret !"* est contesté par des principes de vie privée : *"Ce n'est pas parce que je n'ai rien à cacher que tout doit être public"* [16, Schneier].
- **Vulnérabilités Inhérentes aux Systèmes d'IA et Attaques Spécifiques :**
  L'IA est **alimentée par tout ce que l'humain peut lui donner comme contenu** (documents confidentiels, données médicales, états d'âme), la rendant parfois vulnérable du fait de ses méthodes d'apprentissage.
    1. **Empoisonnement des Données (Data Poisoning Attacks) :** Injection de données corrompues durant l'entraînement pour dégrader la performance ou introduire des backdoors [26, Biggio et al.; 27, Shafahi et al.].
    2. **Attaques Adversariales (Adversarial Attacks / Evasion Attacks) :** Perturbations minimes des entrées en phase d'inférence pour tromper le modèle [28, Szegedy et al.; 20, Goodfellow et al.].
    3. **Extraction de Données et de Modèle (Model Stealing / Extraction) :** Duplication de modèles propriétaires ou reconstruction de données d'entraînement [29, Tramèr et al.; 35, Carlini et al.].
    4. **Inférence d'Appartenance (Membership Inference Attacks) :** Déterminer si un enregistrement spécifique a été utilisé pour l'entraînement [18, Shokri et al.].
    5. **Inversion de Modèle (Model Inversion Attacks) :** Reconstruire des entrées représentatives à partir du modèle [30, Fredrikson et al.].
    6. **Attaques par Détournement de Prompt (Prompt Injection & Jailbreaking) – Principalement pour les LLM :** Instructions spécifiques pour contourner les filtres de sécurité [19, Perez & Ribeiro; 31, Greshake et al.].
- **Limitations Cognitives et Fiabilité Intrinsèque des Modèles :**

    - **Biais Algorithmiques et Sociétaux :** L'IA peut hériter et amplifier les biais présents dans ses données d'entraînement [17, Buolamwini J., Gebru T.].
    - **L'IA "ne sait pas, qu'elle ne sait pas" :** Manque de mécanismes robustes d'estimation de sa propre incertitude.
    - **Hallucinations et Confabulation :** Les modèles génératifs **peuvent "halluciner"** – inventer des informations fausses [21, Ji et al.].
    - **Désir de Répondre et Opacité :** L'IA **souhaite apporter une réponse dans tous les cas**. L'opacité rend difficile la compréhension des décisions [32, Lipton, Z. C.].
- **L'IA comme Outil d'Attaque Cybernétique (Offensive AI) :**
  Utilisation de l'IA pour spear-phishing amélioré, génération de malware polymorphe, cassage de CAPTCHA, etc. [33, Europol].
- **Le Risque Existentiel sur la Confidentialité Numérique (IA et Cryptographie Quantique) :**
  Le risque majeur sur la confidentialité vient de l'augmentation de la puissance de calcul. La convergence de **l'IA et de l'informatique quantique** pourrait **"casser" les algorithmes de chiffrement actuels** (RSA, ECC), menaçant la **fin des secrets numériques** [23, Shor P.W.]. La transition vers la cryptographie post-quantique (PQC) est cruciale [24, NIST].

Conclusion et Perspectives

L'intelligence artificielle est une force de transformation indéniable, dont l'impact sur la cybersécurité est profond et ambivalent. Elle offre des outils d'une puissance inégalée pour renforcer nos défenses, mais introduit simultanément de nouvelles vulnérabilités et peut être instrumentalisée pour des attaques sophistiquées. La démocratisation de l'IA, bien que la cantonnant encore largement au monde virtuel pour ses formes les plus complexes, voit son extension dans le monde physique s'accélérer via des IA spécialisées intégrées dans nos appareils du quotidien. Cette double nature – virtuelle et de plus en plus physiquement incarnée – complexifie le paysage des menaces. La menace de l'informatique quantique sur la cryptographie actuelle, potentiellement accélérée par l'IA, souligne l'urgence d'une adaptation continue.

Naviguer dans ce paysage complexe exige une approche holistique : favoriser l'innovation responsable ("AI for Security") ; développer des contre-mesures robustes ("Security for AI") ; promouvoir la recherche en d'algorithmes post-quantique ; et engager une réflexion éthique et sociétale approfondie. La collaboration internationale, le partage d'informations et la formation continue seront essentiels pour exploiter le potentiel bénéfique de l'IA tout en maîtrisant les risques qu'elle engendre.

Références

1. Asimov, I. (1950). *I, Robot*. Gnome Press.
2. Clarke, A. C. (1968). *2001: A Space Odyssey*. New American Library.
3. Gibson, W. (1984). *Neuromancer*. Ace Books.

4. Cameron, J. (Director). (1984). *The Terminator* [Film]. Orion Pictures.
5. Bengio, Y., Goodfellow, I., & Courville, A. (2016). *Deep Learning*. MIT Press.
6. OpenAI. (2022, November 30). *Introducing ChatGPT*. OpenAI Blog.
7. Vaswani, A., et al. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
8. Hugging Face. (n.d.). *About Us*.
9. Touvron, H., et al. (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv:2307.09288*.
10. BigScience Workshop. (2022). BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *arXiv:2211.05100*.
11. Li, Y., et al. (2022). Competition-Level Code Generation with AlphaCode. *Science*, 378(6624), 1092-1097.
12. SANS Institute. (Year). *AI and Machine Learning in Cybersecurity: A SANS Survey*. [Lien ou titre complet]
13. IBM Security. (n.d.). *AI for a Smarter, Faster Cyber Defense*. [Lien ou titre complet]
14. Darktrace. (n.d.). *Self-Learning AI Technology*. [Lien vers le site]
15. Vectra AI. (n.d.). *Attack Signal Intelligence*. [Lien vers le site]
16. Schneier, B. (2015). *Data and Goliath: The Hidden Battles to Collect Your Data and Control Your World*. W. W. Norton & Company.
17. Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Conference on fairness, accountability and transparency*, 77-91.
18. Shokri, R., et al. (2017). Membership inference attacks against machine learning models. *IEEE S&P*.
19. Perez, F., & Ribeiro, M. T. (2022). Ignore Previous Prompt: Attack Techniques For Language Models. *arXiv:2211.09527*.
20. Goodfellow, I. J., et al. (2014). Explaining and harnessing adversarial examples. *arXiv:1412.6572*.
21. Ji, Z., et al. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12), 1-38.
22. Seymour, J., & Tully, P. (2017). Weaponizing data science for social engineering. *Black Hat USA*.
23. Shor, P. W. (1994). Algorithms for quantum computation. *IEEE FOCS*.
24. NIST. (n.d.). *Post-Quantum Cryptography Program*. [Lien vers le site NIST]
25. Foucault, M. (1975). *Surveiller et punir : Naissance de la prison*. Gallimard.
26. Biggio, B., et al. (2012). Poisoning attacks against support vector machines. *ICML*.

27. Shafahi, A., et al. (2018). Poison frogs! Targeted clean-label poisoning attacks on neural networks. *NeurIPS*.

28. Szegedy, C., et al. (2013). Intriguing properties of neural networks. *arXiv:1312.6199*.

29. Tramèr, F., et al. (2016). Stealing machine learning models via prediction apis. *USENIX Security*.

30. Fredrikson, M., et al. (2015). Model inversion attacks that exploit confidence information. *ACM CCS*.

31. Greshake, K., et al. (2023). Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. *arXiv:2302.12173*.

32. Lipton, Z. C. (2018). The mythos of model interpretability. *Queue*, 16(3), 31-57.

33. Europol. (2020). *Malware, Hacking and Criminal Use of AI*. Europol Innovation Lab.

34. Orwell, G. (1949). *Nineteen Eighty-Four*. Secker & Warburg.

35. Carlini, N., et al. (2021). Extracting training data from large language models. *USENIX Security*.

36. Microsoft. (2023). [Titre de l'annonce de Copilot dans Windows 11]. *Microsoft Official Blog*. [Lien spécifique]

37. [Auteur/Organisation]. (Année). [Titre de l'article sur les NPUs]. *Publication Technique/Site d'Analyste*. [Lien spécifique]

38. Apple Inc. (n.d.). [Titre de la documentation sur le Neural Engine]. *Apple Developer Documentation*. [Lien spécifique]

39. Google. (Année). [Titre de l'article de blog sur les fonctionnalités IA d'Android]. *Google AI Blog / Android Developers Blog*. [Lien spécifique]

40. Google. (n.d.). [Titre de la documentation sur Google Tensor]. *Google Store / AI Blog*. [Lien spécifique]