Vincent GENOT

L'Intelligence Artificielle et cybersécurité

Chapitre 3 : Le cybercontrôle

De la surveillance historique au cybercontrôle cognitif et à la nécessaire résilience humaine dans un monde numérique omniprésent

Université BORDEAUX MONTAIGNE
Cyber criminalités

Promotion 2024-2025

Nous avons vu dans le Second chapitre, que l'intelligence artificielle n'est un qu'un programme, une fonctionnalité étendue, ou un système d'information à part entière qui peut être victime et la cible, comme toute entité ou actif numérique, de risques cyber. Nous allons maintenant dans ce nouveau chapitre explorer ces menaces, et les risques d'une vulgarisation de l'IA dans nos outils de tous les jours avec une approche « **orwellienne** » pour mieux nous préparer à ce futur tant craint et prophétisé et qui semble indéfectiblement s'annoncer.

Résumé : L'Intelligence Artificielle (IA) reconfigure radicalement le paysage de la cybersécurité, offrant des capacités défensives inédites tout en introduisant des vecteurs d'attaque d'une complexité nouvelle et d'une portée sans précédent. Cet article examine en profondeur cette dualité fondamentale. Il explore de manière exhaustive les menaces émergentes induites par l'IA, incluant les attaques sophistiquées contre ses propres infrastructures (empoisonnement de données, attaques par d'autres ia, vol de modèles), le détournement d'IA intégrées dans nos vies quotidiennes, et l'amplification systémique de la désinformation à une échelle industrielle.

S'inscrivant dans la lignée des travaux fondateurs de **Mattelart et Vitalis (2014)** sur l'évolution historique du profilage des populations, du "livret ouvrier" au "cybercontrôle", cet article analyse comment l'IA, en tant que technologie de pouvoir, exacerbe et automatise les logiques de surveillance, de classification et de contrôle social.

Une attention particulière et approfondie est accordée à l'impact insidieux et potentiellement dévastateur de l'IA sur les fonctions cognitives humaines – érodant l'esprit critique, la capacité de raisonnement autonome, la mémoire, l'attention et le libre arbitre – et à l'émergence d'une **neurocybercriminalité** qui cible délibérément nos vulnérabilités neuropsychologiques. S'appuyant sur les recherches de Donzel (2025) sur l'ingénierie cognitive et ses manifestations dans l'espace numérique, de Darcy et al. (2025) sur les mécanismes cognitifs de la désinformation et les stratégies de résilience, et des Docteurs Teboul et Malbos (2023) sur la neuropsychologie de la cybercriminalité ("Dark Cyber").

Cet article souligne l'urgence de développer des mesures de sécurité robustes et adaptatives, des cadres éthiques et légaux exigeants, et une approche sociétale humanocentrique pour comprendre, anticiper et atténuer ces risques multidimensionnels complexes de l'arrivée de l'IA dans notre vie de tous les jours pour les générations futures.

Mots-clés : Intelligence Artificielle (IA), Cybersécurité, Profilage des Populations, Cybercontrôle, Gouvernementalité Algorithmique, Ingénierie Cognitive, Neurocybercriminalité, Dark Cyber, Cyberpsychologie, Désinformation, Vulnérabilité Humaine, Fonctions Cérébrales, Biais Cognitifs, Manipulation Comportementale, Éthique de l'IA, Écologie Informationnelle, Résilience Cognitive, Surveillance de Masse, Capitalisme de Surveillance, Attention Economy.

1. Introduction : Du Livret Ouvrier au Cyber contrôle Cognitif Propulsé par l'IA

L'avènement de l'Intelligence Artificielle (IA) constitue une rupture technologique majeure comparable aux grandes révolutions industrielles, elle redessine les contours de la cybersécurité avec une ambivalence fondamentale. D'un côté, l'IA offre des promesses considérables pour renforcer nos défenses : capacités prédictives accrues, détection d'anomalies en temps réel à grande échelle, automatisation des réponses aux incidents, et analyse de volumes de données de sécurité (logs, flux réseau, renseignements sur les menaces) dépassant largement les capacités humaines (Liao et al., 2023 ; Apruzzese et al., 2023). De l'autre, elle dote les acteurs malveillants – qu'ils soient étatiques, criminels ou idéologiques – d'outils d'une puissance et d'une sophistication inédites, abaissant les barrières à l'entrée pour des attaques complexes et élargissant considérablement la surface d'attaque des systèmes d'information et des sociétés elles-mêmes.

Au-delà de ces menaces techniques directes, cet article se propose d'explorer une dimension plus insidieuse et potentiellement plus déstabilisatrice : l'influence de l'IA sur la cognition humaine et son rôle central dans l'amplification et la transformation des mécanismes de contrôle social. Pour comprendre cette dynamique, il est essentiel de l'inscrire dans une perspective historique. **Armand Mattelart et André Vitalis, dans leur ouvrage séminal "Le profilage des populations : Du livret ouvrier au cybercontrôle" (2014)**, ont brillamment démontré que les techniques de fichage, de surveillance, de classification et de catégorisation des individus ne sont pas une invention récente. Le livret ouvrier du XIXe siècle en France, par exemple, n'était pas un simple document administratif ; il incarnait un instrument de contrôle de la mobilité, de la qualification professionnelle et, implicitement, de la moralité et de la docilité de la classe laborieuse. C'était une technologie de pouvoir visant à discipliner et à gérer les populations considérées comme potentiellement dangereuses ou instables. Mattelart et Vitalis tracent ainsi une généalogie des dispositifs de contrôle, montrant comment les logiques de surveillance étatique et, de plus en plus, privée, se sont affinées avec les avancées technologiques, de la biométrie naissante aux premières bases de données informatisées.

Le concept de "cybercontrôle" qu'ils développent, désigne la sophistication, l'extension et, de manière cruciale, la privatisation de ces logiques par les technologies numériques et les nouveaux acteurs économiques du capitalisme informationnel. L'IA, avec ses capacités exponentielles d'analyse de données massives (Big Data), d'apprentissage automatique (Machine Learning) permettant de déduire des corrélations et des schémas comportementaux complexes, et de profilage prédictif, porte ce cybercontrôle à un niveau de granularité, d'omniprésence et d'efficacité sans précédent. Ce n'est plus seulement la surveillance des actions passées, mais la prédiction et la modulation des comportements futurs qui sont en jeu. Ce "capitalisme de surveillance", brillamment analysé par Shoshana Zuboff (2019), repose sur l'extraction et la marchandisation des données comportementales comme nouvelle matière première.

Dans ce contexte, l'IA n'est pas un simple outil neutre ; elle devient le moteur d'un cyber contrôle qui affecte non seulement nos libertés civiles et notre vie privée, mais aussi, et c'est une thèse centrale de cet article, nos fonctions cérébrales et nos processus cognitifs fondamentaux. Comme le souligne Donzel (2025) dans ses travaux sur l'ingénierie cognitive, nos facultés mentales sont de plus en plus « à la merci des cyber-attaquants » [OCR Donzel p.2] et des systèmes algorithmiques qui façonnent notre environnement informationnel. Parallèlement, le phénomène de la désinformation, exacerbé par l'IA générative et les algorithmes de diffusion, est perçu comme un risque sociétal majeur (Darcy et al., 2025, Abstract, p.2). Ce rapport de l'Institut

Jean Nicod met en lumière les limites des approches actuelles, souvent réactives et basées sur un modèle simpliste de la psychologie humaine, pour contrer un phénomène qui prospère sur des "vulnérabilités multiples – cognitives, affectives, sociales, économiques et institutionnelles – et sur un écosystème informationnel profondément dégradé". Les travaux des Docteurs Teboul et Malbos (2023) dans "Dark Cyber" vont plus loin en identifiant l'émergence d'une **neurocybercriminalité**, qui cible délibérément et scientifiquement nos vulnérabilités neuropsychologiques.

Ce chapitre se propose donc de synthétiser et d'articuler ces différentes perspectives pour offrir une compréhension holistique des enjeux de cybersécurité à l'ère de l'IA. Il s'agira d'explorer comment l'IA est utilisée pour perfectionner les techniques de profilage et de cybercontrôle, comment cela crée de nouvelles vulnérabilités systémiques et individuelles, et surtout, comment cette nouvelle donne technologique impacte notre cognition, notre capacité à prendre des décisions éclairées et à exercer notre libre arbitre. Enfin, nous esquisserons les pistes d'une résilience collective et individuelle, impliquant des mesures techniques, éthiques, légales, éducatives et sociétales.

2. L'IA comme Vecteur de Vulnérabilités Cybernétiques : Le Profilage, la Manipulation et la Dégradation Systémique au Cœur des Menaces

L'intégration de l'IA dans la quasi-totalité des systèmes numériques crée et amplifie un large éventail de vulnérabilités, où le profilage des populations et la manipulation cognitive jouent un rôle de plus en plus central et préoccupant.

- **2.1. Surveillance Accrue, Profilage Généralisé et Érosion de la Confidentialité : L'Ère du Cybercontrôle Omniprésent**
  Les systèmes de surveillance alimentés par l'IA, initialement présentés comme des outils au service de la sécurité publique (par exemple, la lutte contre le terrorisme ou la criminalité via la reconnaissance faciale dans les espaces publics, l'analyse prédictive des zones à risque) ou de l'optimisation des services (gestion des flux urbains, personnalisation des offres commerciales), réalisent de facto le "cyber contrôle" à une échelle massive et intrusive, tel qu'anticipé par Mattelart et Vitalis. Le profilage constant des citoyens, effectué par une myriade d'acteurs étatiques et, de manière encore plus pervasive, par des entreprises privées (géants du web, courtiers en données, annonceurs), crée des "fiches numériques" ou "doubles numériques" d'une granularité et d'une profondeur inégalées. Ces profils ne se contentent plus d'agréger des données déclaratives (nom, âge, adresse) mais intègrent des informations comportementales (historique de navigation, achats, déplacements, interactions sociales en ligne), des données biométriques (reconnaissance faciale, vocale), des inférences psychométriques (traits de personnalité, opinions politiques, orientation sexuelle, état émotionnel) et même des prédictions sur les actions futures (probabilité de contracter une maladie, de commettre un délit, de quitter un emploi, etc.).
  Cette accumulation de données et leur analyse par l'IA ne servent pas seulement à une surveillance rétrospective, mais alimentent une "gouvernementalité algorithmique" (Rouvroy & Berns, 2013). **Ce concept désigne un mode de gouvernement où les décisions concernant les individus et les collectifs sont de plus en plus déléguées à des systèmes algorithmiques**, souvent opaques ("boîtes noires") et fonctionnant sur la base de corrélations statistiques plutôt que de causalités établies ou de jugements individualisés (Netflix, Prime, Deezer qui vous proposent des choix !).

L'accès à des services essentiels (crédit, assurance, emploi, éducation, justice), la tarification dynamique, l'évaluation des risques, l'attribution de "scores sociaux" (comme le système de crédit social en Chine, mais avec des manifestations plus diffuses en Occident) peuvent ainsi être modulés par des profils algorithmiques, créant de nouvelles formes d'inégalités, de discriminations (biais algorithmiques reproduisant ou amplifiant les préjugés sociétaux) et d'exclusion sociale, tout en réduisant la capacité des individus à comprendre et contester les décisions qui les affectent. La perte de confidentialité n'est alors plus seulement un risque d'exposition de secrets personnels, mais une condition de participation à la vie sociale et économique, où chaque action est susceptible d'être tracée, analysée et intégrée à un profil global (Andrejevic, 2007).

- **2.2. Attaques contre les Infrastructures de l'IA et les Données de Profilage : Le Trésor Convoité des "Data Lakes"**
  Les vastes entrepôts de données ("data lakes") contenant ces profils individuels et collectifs constituent des cibles de très haute valeur pour les cybercriminels et les acteurs étatiques malveillants. Les attaques contre les serveurs d'IA, les bases de données des plateformes numériques (réseaux sociaux, sites de e-commerce, services de santé, institutions financières), ou les systèmes des courtiers en données (data brokers) peuvent mener à des fuites de données d'une ampleur et d'une sensibilité considérables. Des exemples récents de fuites de données massives en France : Free, Sport 2000,... les nombreuses violations de données dans le secteur de la santé ou des services publics..., sur les broker des données de carte de fidélités, ou encore chez Ledger (affectant les détenteurs de cryptomonnaies), illustrent la vulnérabilité de ces "gisements" informationnels.

  Les conséquences de telles fuites sont multiples :

  - **Usurpation d'identité sophistiquée :** La richesse des données compromises permet de créer des usurpations d'identité beaucoup plus crédibles et difficiles à détecter, allant au-delà du simple vol de numéros de carte de crédit pour inclure la prise de contrôle de comptes en ligne, la création de faux profils, voire la manipulation de processus administratifs ou légaux.
  - **Chantage et extorsion ciblés :** Des informations intimes ou compromettantes issues des profils peuvent être utilisées pour exercer un chantage sur des individus ou des organisations.
  - **Réutilisation à des fins de surveillance étatique ou d'espionnage industriel :** Des acteurs étatiques peuvent exploiter ces données pour des opérations de renseignement, de surveillance de dissidents, ou d'espionnage économique.
  - **Manipulation politique et sociale à grande échelle :** Les profils dérobés peuvent alimenter des campagnes de désinformation ou d'influence ciblées, comme l'a illustré le scandale Cambridge Analytica (Cadwalladr & Graham-Harrison, 2018).

  La sécurisation de ces infrastructures et des données qu'elles contiennent est donc un **enjeu majeur**, et elle va se complexifier par la nature distribuée des agents IA et la nécessité d'accéder à de grands volumes de données pour l'entraînement des

modèles. Les vulnérabilités des serveurs de contrôle des modèles d'IA (Model Control Plane - MCP), qui gèrent les interactions entre les modèles et les applications, représentent un point d'entrée critique pour les attaquants.

- **2.3. Détournement d'IA Intégrées Exfiltrer des Données Sensibles, Manipuler les Interactions et pour Affiner le Cybercontrôle :**
  Les IA ne sont plus cantonnées à des serveurs distants mais sont de plus en plus intégrées dans nos appareils personnels (smartphones avec assistants vocaux, ordinateurs portables, objets connectés) et nos applications quotidiennes (logiciels bureautiques, navigateurs web, applications de messagerie). **Ces IA "embarquées"** ou "locales" deviennent des **agents de collecte de données** en continu, **affinant** en permanence les **profils utilisateurs pour une personnalisation accrue des services**, mais aussi pour une potentielle surveillance, *sans tomber dans la paranoïa*, mais au vu des profils personnels crées par l'ia qui vont démultiplier ces prochains mois (années), leur détournement ouvre la voie à de nouvelles formes d'attaques :

  - **Exfiltration de données contextuelles et intimes :** Une IA compromise sur un appareil personnel peut accéder à un flux continu de données sensibles : conversations audios, frappes clavier, historique de navigation, géolocalisation, données biométriques issues de capteurs, contenu des emails et des documents. *Des vulnérabilités dans des IA comme Microsoft Copilot, permettant par exemple d'accéder à des mots de passe stockés localement ou d'exfiltrer des données depuis des applications comme SharePoint via des techniques de "prompt injection" ou d'exploitation de permissions excessives, illustrent ce risque.*
  - **Manipulation des interactions homme-machine :** Une IA détournée peut subtilement modifier les informations présentées à l'utilisateur, l'orienter vers des sites malveillants, lui faire approuver des transactions frauduleuses, ou encore enregistrer ses identifiants de connexion.
  - **Création de "chevaux de Troie" intelligents :** Des IA malveillantes pourraient être conçues pour opérer discrètement sur un appareil pendant de longues périodes, collectant des informations et attendant un signal pour déclencher une action nuisible (exfiltration de données, rançongiciel, participation à un botnet). La prolifération des objets connectés (IoT), souvent dotés de capacités d'IA rudimentaires et de faibles niveaux de sécurité, multiplie les points d'entrée potentiels pour ce type de cyber contrôle localisé et d'exfiltration de données.

  Cela ouvre un **nouveau champ des « Possibles »** pour les acteurs malveillants.

- **2.4. Désinformation Hyper-Ciblée et Ingénierie Sociale Augmentée grâce au Profilage Algorithmique alimentée par IA.**

  L'IA générative (comme les modèles GPT-3/4, DALL-E, Midjourney) a révolutionné la capacité à créer des contenus textuels, visuels et audio d'un réalisme saisissant à très faible coût. Combinée au profilage algorithmique détaillé, elle permet de concevoir et de

diffuser des campagnes de désinformation et d'ingénierie sociale d'une efficacité redoutable.

- **Personnalisation des narratifs trompeurs :** Au lieu de diffuser un message unique à une large audience, l'IA peut adapter le contenu, le ton, le style et même la langue d'un message de désinformation aux caractéristiques psychologiques, aux croyances préexistantes, aux affiliations politiques et aux vulnérabilités émotionnelles de chaque individu ou micro-segment de population (Darcy et al., 2025, p.17 sur les facteurs identitaires). Par exemple, *une fausse information sur un vaccin pourra être présentée avec des arguments pseudo-scientifiques à une audience éduquée mais méfiante, et avec des témoignages émotionnels à une audience plus sensible à ce type de récits*.

- **Création de faux profils et d'agents conversationnels malveillants (Social Bots) :** L'IA permet de générer des profils de réseaux sociaux (photos, biographies, historiques de posts) extrêmement crédibles et difficiles à distinguer de comptes humains authentiques. Ces "deepfake identities" peuvent être utilisées pour infiltrer des communautés en ligne, propager de la désinformation, amplifier artificiellement certains messages (astroturfing), ou mener des opérations d'ingénierie sociale ciblées (par exemple, des arnaques sentimentales ou des tentatives de spear-phishing). Les chatbots malveillants peuvent engager des conversations personnalisées pour extraire des informations ou influencer des opinions (ping @yoni et son article de DU).

- **Automatisation et industrialisation de la désinformation :** L'IA permet d'automatiser la création et la diffusion de contenus trompeurs à une échelle industrielle, submergeant les plateformes et les capacités de vérification humaine (fact-checking). La rapidité de propagation et la capacité d'adaptation des campagnes de désinformation augmentées par l'IA rendent les contre-mesures traditionnelles de plus en plus difficiles à mettre en œuvre efficacement. Les travaux de Mattelart et Vitalis sur la manière dont le profilage historique servait à "catégoriser pour mieux régner" ou influencer trouvent ici un écho puissant ***: l'IA devient l'outil par excellence pour une segmentation fine et une influence comportementale à grande échelle***, dans des nouveaux Datacenters au plus près de la population ?

- **2.5. L'IA dans les Opérations Cyber Offensives et la Neurocybercriminalité Ciblée : Vers une Exploitation Psychologique Précise**
  Les acteurs malveillants, qu'ils soient criminels ou étatiques, intègrent de plus en plus l'IA dans leurs méthodologies offensives, allant au-delà de la simple automatisation pour s'orienter vers une exploitation psychologique précise.

  - **Identification automatisée de vulnérabilités logicielles et humaines :** L'IA peut être entraînée pour scanner des codes sources et identifier des failles de sécurité (0-day) ou pour analyser des comportements en ligne et des communications afin de repérer les individus les plus susceptibles de cliquer sur un lien de phishing, de divulguer des identifiants, ou d'être sensibles à une manipulation émotionnelle. Le profilage avancé par l'IA permet de construire des "modèles de vulnérabilité" individuels.

- **Développement de malware adaptatif et évasif :** L'IA peut contribuer à créer des logiciels malveillants capables de modifier leur comportement pour échapper aux détections antivirus (malware polymorphe ou métamorphique), d'adapter leurs charges utiles en fonction de la cible, ou d'apprendre des environnements dans lesquels ils sont déployés pour optimiser leur persistance et leur efficacité. (Création de Worm GPT, par des versions détournées de LLM Mistral AI et Grok)
- **Neurocybercriminalité :** Ce concept, développé par les Docteurs Teboul et Malbos (2023) dans "Dark Cyber", désigne l'utilisation délibérée des connaissances en neurosciences et en psychologie cognitive, augmentée par l'IA, pour concevoir des cyberattaques qui exploitent directement les failles intrinsèques du fonctionnement cérébral humain. Cela inclut :
  - La surcharge cognitive délibérée pour réduire la vigilance.
  - L'exploitation des biais cognitifs (ancrage, confirmation, autorité, etc.) pour induire des erreurs de jugement.
  - La manipulation émotionnelle (peur, urgence, curiosité, empathie) pour court-circuiter le raisonnement rationnel.
  - L'utilisation de techniques de conditionnement et de persuasion subliminale via des interfaces numériques.

  => Les "Dark LLM" (Modèles de Langage Larges débridés ou spécialisés dans la génération de contenus malveillants) deviennent des outils clés pour la neurocybercriminalité, permettant de générer des leurres (phishing, vishing, smishing) d'une crédibilité et d'une personnalisation extrêmes.

- **2.6. Prolifération des Bots Malveillants mérique et Dégradation de l'Écosystème Informationnel : Vers une "Pollution Cognitive" de l'Espace Numérique**
  L'IA démultiplie la capacité à créer et opérer des armées de bots malveillants. Selon certaines études (par exemple, les rapports annuels de sociétés comme Imperva sur le trafic des bots), le trafic internet généré par les bots (bons et mauvais) dépasse déjà celui généré par les humains.
  - **Automatisation des attaques :** Les bots dopés à l'IA peuvent mener des attaques DDoS plus sophistiquées et adaptatives, des campagnes de spam et de phishing à grande échelle, et des tentatives de "credential stuffing" (test massif d'identifiants volés sur de multiples plateformes).
  - **Manipulation de l'opinion et désinformation :** Ils sont massivement utilisés pour créer de fausses tendances sur les réseaux sociaux (trending topics), amplifier artificiellement certains messages (astroturfing), discréditer des opposants, ou encore inonder les sections de commentaires de propagande ou de discours haineux.
  - **Automatisation des pseudo-sites d'information :** De faux sites de médias automatisés à l'IA sont de plus en plus visibles en ligne. Boostés par les algorithmes de Google, ils menacent la réputation des acteurs traditionnels du secteur et amplifient la désinformation.

- **Érosion de la confiance et "pollution cognitive" :** Cette omniprésence de contenus et d'interactions non authentiques contribue à la "dégradation globale de l'écosystème informationnel" (Darcy et al., 2025, p.4). Elle rend plus difficile pour les utilisateurs de distinguer le vrai du faux, l'humain de l'artificiel, et peut conduire à une forme de "pollution cognitive" où la méfiance devient la norme et où la capacité à s'engager dans un débat public sain est compromise. On peut même envisager, comme une hypothèse extrême, la constitution d'une "entité agglomérée" de ces intelligences artificielles distribuées, agissant de manière chaotique ou coordonnée pour déstabiliser les réseaux et les sociétés dans une vision dystopique extrême.

.

3. La Menace Insidieuse : Ingénierie Cognitive, Désinformation et l'Ère de la Neurocybercriminalité sous le Prisme du Cybercontrôle Algorithmique

Au-delà des vulnérabilités techniques et des attaques directes contre les systèmes comme présenté au chapitre 2, le risque peut-être plus profond, plus diffus et plus durable suite au déploiement généralisé de l'IA qui réside dans son impact sur les fonctions cognitives humaines et dans sa capacité à remodeler notre perception du monde, nos processus de décision et, in fine, notre autonomie. Cette dimension s'inscrit dans la continuité historique des mécanismes de contrôle social analysés par Mattelart et Vitalis, mais elle acquiert avec l'IA une puissance et une subtilité inédites. Le "cyber contrôle" ne se limite plus à une surveillance externe de nos actions, mais s'étend à une potentielle **internalisation des contraintes et des influences par le biais d'une ingénierie cognitive sophistiquée**.

- **3.1. Ingénierie Cognitive et Neurocybercriminalité : Le Profilage Détaillé comme Prérequis à la Manipulation Précise de l'Esprit**
  L'"ingénierie cognitive", telle que conceptualisée par Donzel (2025) [OCR Donzel p.5, 8-10], désigne l'ensemble des techniques visant à comprendre, modéliser, influencer, voire altérer les processus mentaux humains. Lorsqu'elle est employée à des fins malveillantes, elle devient le fondement de la **neurocybercriminalité**, un concept exploré par les Docteurs Teboul et Malbos (2023) dans "Dark Cyber". Ces derniers soulignent que les cybercriminels ne se contentent plus d'exploiter des failles logicielles, mais ciblent délibérément les "vulnérabilités neuropsychologiques de l'esprit humain comme les biais cognitifs, les émotions et les types de personnalités".
  Le **profilage détaillé des individus**, rendu possible par la collecte massive de données et leur analyse par l'IA (Mattelart & Vitalis, 2014 ; Zuboff, 2019), est le prérequis essentiel à cette forme d'ingénierie cognitive malveillante. Connaître les schémas de pensée, les anxiétés, les valeurs, les affiliations groupales, les habitudes de consommation d'information et les traits de personnalité d'un individu ou d'un groupe permet de :

  - **Adapter les techniques de manipulation :** Un message de phishing, une campagne de désinformation ou une tentative d'escroquerie peuvent être finement calibrés pour exploiter les biais cognitifs spécifiques d'une cible (biais d'autorité, de confirmation, d'ancrage, de rareté, etc.), ses déclencheurs émotionnels (peur, cupidité, curiosité, empathie, colère, sentiment d'injustice), ou ses vulnérabilités psychologiques identifiées (faible estime de soi, besoin de reconnaissance, anxiété).

- **Augmenter la crédibilité des leurres :** L'IA générative peut créer des messages, des images ou des profils synthétiques qui "sonnent vrai" pour la cible parce qu'ils sont alignés avec ses attentes, son langage, ses centres d'intérêt, ou les caractéristiques de son réseau social.

- **Identifier les "maillons faibles" :** L'analyse prédictive peut servir à identifier les individus les plus susceptibles de succomber à une manipulation, de relayer une fausse information, ou de devenir des vecteurs involontaires d'une attaque (par exemple, en cliquant sur un lien malveillant).
  Darcy et al. (2025, p.16-17) corroborent cette analyse en identifiant une panoplie de facteurs individuels (sociodémographiques, traits de personnalité comme ceux de la "Triade Sombre" – narcissisme, machiavélisme, psychopathie –, vulnérabilités cognitives et épistémiques, facteurs émotionnels et existentiels, facteurs identitaires et de polarisation) qui augmentent la réceptivité à la désinformation et, par extension, à d'autres formes de manipulation cognitive.
  Le "piratage cognitif" [Jourt Pineau, 2024], c'est-à-dire la prise de contrôle des processus de pensée de la victime, devient une réalité tangible.

- **3.2. Cyberdépendance, Environnements Numériques Immersifs et Affaiblissement des Défenses Cognitives dans un Cadre de Cybercontrôle Normalisé :** L'omniprésence des technologies numériques, souvent conçues pour être addictives (Alter, 2017), et l'immersion croissante dans des environnements virtuels ou augmentés, façonnent nos capacités cognitives et peuvent les affaiblir, nous rendant plus vulnérables au cyber contrôle.

  - **3.2.1. Surcharge Informationnelle Chronique et Baisse de l'Esprit Critique :** L'exposition constante à un flux ininterrompu d'informations, de notifications et de sollicitations via les smartphones et les réseaux sociaux conduit à une **surcharge cognitive chronique**. Notre cerveau, comme l'explique Gérald Bronner (2021) dans "L'Apocalypse cognitive", est biologiquement programmé pour prêter attention à la nouveauté et aux signaux saillants, mais il est mal équipé pour traiter la surabondance informationnelle actuelle. Cela se traduit par une attention fragmentée, une réduction de la capacité de concentration profonde (deep work, Cal Newport, 2016), une lecture en diagonale, et une diminution de la propension à l'analyse critique et à la vérification des sources (Darcy et al., 2025, p.4). L'"économie de l'attention" (Citton, 2014 ; [OCR Donzel p.23]) monétise cette fragmentation, les plateformes étant incitées à maximiser le temps d'engagement plutôt que la qualité de l'information ou le bien-être cognitif de l'utilisateur. Dans ce contexte, les interventions mal calibrées, comme un débunking maladroit, peuvent même être contre-productives en renforçant la familiarité avec une fausse information ou en aggravant la défiance (Darcy et al., 2025, p.5).

  - **3.2.2. Délestage Cognitif (Cognitive Offloading) et Atrophie des Compétences Mentales :** La facilité avec laquelle nous pouvons déléguer des tâches cognitives à nos appareils (mémoire via les moteurs de recherche – "l'effet Google" de Sparrow et al., 2011 [OCR Donzel p.25] ; orientation via le GPS ; calcul via les calculatrices intégrées ; rédaction via les IA génératives) a des conséquences sur nos propres capacités. Si ce "délestage cognitif" [OCR Donzel p.24-25] peut libérer des ressources mentales pour d'autres tâches, MAIS une dépendance excessive peut entraîner une **atrophie des compétences cognitives sous-**

jacentes (mémoire de travail, mémoire à long terme, capacités d'orientation spatiale, raisonnement analytique) (Carr, 2008). Cette érosion progressive de nos facultés peut nous rendre plus dépendants des systèmes algorithmiques pour prendre des décisions, et donc plus vulnérables à leur influence ou à leur manipulation.

- **3.2.3. L'Ère de la Post-Vérité : Quand le Profilage Algorithmique Nourrit les Bulles de Filtres et la Polarisation Cognitive :** Le contenu généré par l'IA et les algorithmes de recommandation des plateformes (basés sur le profilage intensif de nos comportements, préférences et interactions) contribuent massivement à la création et au renforcement des "bulles de filtres" (Pariser, 2011 [OCR Donzel p.40]) et des "chambres d'écho" idéologiques. En nous exposant prioritairement à des informations qui confirment nos croyances préexistantes (biais de confirmation) et en nous isolant des perspectives divergentes, ces mécanismes algorithmiques favorisent la polarisation des opinions, l'extrémisation des points de vue, et une perception biaisée de la réalité (Sunstein, 2017).

  L'IA peut ainsi fabriquer des "réalités sur mesure" pour chaque utilisateur, rendant le consensus social et la délibération démocratique de plus en plus difficiles. Darcy et al. (2025, p.19) soulignent que la "crise de l'information" est profondément ancrée dans des facteurs sociaux et institutionnels, incluant la "dégradation de la confiance institutionnelle" (p.20), que ces bulles informationnelles peuvent exacerber.

- **3.2.4. Cyberdépendance, Santé Mentale et Vulnérabilité Accrue :** Les travaux d'Eric Malbos [2, 3-doc2, 4-doc2, 5-doc2], notamment sur l'utilisation de la Réalité Virtuelle (VR) pour traiter les addictions (jeu, substances, etc.) par exposition contrôlée et désensibilisation, mettent en lumière, par contraste, les puissants mécanismes de dépendance que les technologies numériques immersives et engageantes peuvent induire. La conception de nombreux jeux vidéo, réseaux sociaux et applications repose sur des principes de "design persuasif" et de "gamification" qui exploitent les circuits de la récompense dans notre cerveau (système dopaminergique), pouvant mener à une utilisation compulsive et à une véritable **cyberdépendance** (Griffiths, 2005). Cette dépendance peut entraîner des conséquences délétères sur la santé mentale (anxiété, dépression, troubles du sommeil, isolement social) et, par ricochet, augmenter la vulnérabilité aux manipulations en ligne, les individus en situation de détresse psychologique étant souvent des cibles plus faciles.

- **3.3. Le Rôle Central de l'IA dans la Conception et la Diffusion de Manipulations Comportementales Hyper-Personnalisées :**
  Les techniques de manipulation comportementale, déjà présentes dans le marketing et le design d'interface, sont décuplées par les capacités de profilage et de personnalisation de l'IA.

  - **3.3.1. Nudges, Dark Patterns et Boucles Addictives Optimisés par l'IA :** Les "nudges" (Thaler & Sunstein, 2008), ces incitations douces conçues pour orienter les choix sans contraindre, peuvent être optimisés par l'IA pour une efficacité maximale sur des segments de population spécifiques. Plus problématiques, les "dark patterns" – ces interfaces utilisateurs délibérément trompeuses conçues

pour amener l'utilisateur à effectuer des actions non souhaitées (achats non désirés, partage de données excessif, abonnements cachés) – peuvent être personnalisés dynamiquement par l'IA en fonction du profil de vulnérabilité de chaque utilisateur (Brignull, n.d.). Les boucles addictives, basées sur des cycles de déclencheur-action-récompense variable-investissement (Eyal, 2014), sont affinées par l'IA pour maximiser l'engagement et la rétention, parfois au détriment du bien-être de l'utilisateur. Darcy et al. (2025, p.24) évoquent les "nudges comportementaux" (Levier 4) comme un levier possible pour contrer la diffusion impulsive, mais leur efficacité dépend de leur conception éthique et de leur acceptabilité.

- **3.3.2. Exploitation Fine des États Émotionnels et des Vulnérabilités Psychologiques Identifiés par Profilage IA :** L'IA peut analyser en temps réel des signaux comportementaux (ton de la voix, expressions faciales via webcam, rythme de frappe, contenu des messages) pour inférer l'état émotionnel d'un utilisateur. Ces informations peuvent ensuite être exploitées pour présenter des contenus ou des sollicitations au moment précis où la personne est la plus vulnérable ou la plus réceptive (par exemple, une publicité pour un produit réconfortant en période de stress, ou une sollicitation financière frauduleuse ciblant une personne identifiée comme anxieuse ou isolée). Le scandale Cambridge Analytica, où les profils psychométriques de millions d'utilisateurs de Facebook ont été utilisés pour des campagnes d'influence politique ciblées, a été un révélateur de ces pratiques (Isaak & Hanna, 2018). "Dark Cyber" (Teboul & Malbos, 2023) insiste sur le rôle crucial des émotions comme porte d'entrée pour les cyberattaques par ingénierie cognitive.

4. Impact sur les Générations Futures : Risques de Normalisation du Cybercontrôle, d'Atrophie des Capacités Cognitives Essentielles et d'Évolution de la Pensée Critique : Les "natifs numériques", c'est-à-dire les générations qui ont grandi avec une omniprésence des technologies numériques et de l'IA, sont confrontés à des défis spécifiques qui pourraient façonner durablement leur développement cognitif et leur rapport au monde.

- **4.1. Normalisation de la surveillance et du profilage :** Grandir dans un environnement où la collecte de données et le profilage sont des pratiques courantes et souvent invisibles peut conduire à une **normalisation de la surveillance** et à une moindre sensibilité aux enjeux de vie privée et d'autonomie (Nissenbaum, 2009, sur l'intégrité contextuelle). Le "cyber contrôle" risque d'être perçu non pas comme une contrainte, mais comme une composante "naturelle" de l'environnement numérique, avec une potentielle érosion de l'attente de confidentialité et du désir d'espaces non surveillés.
- **4.2. Impact sur le développement cognitif et la pensée critique :** Une exposition précoce et intensive à des environnements numériques hyper-stimulants, à des flux d'information fragmentés, et à des interactions sociales médiatisées par des algorithmes peut avoir des conséquences sur le développement des fonctions cognitives supérieures :
  - **Attention et concentration :** Difficulté à maintenir une attention soutenue, préférence pour le multitâche (qui s'avère souvent moins efficace), et une plus grande distractibilité (Ophir et al., 2009). L'omniprésence d'outils d'IA conçus pour fournir des réponses rapides pourrait réduire la patience cognitive nécessaire pour s'attaquer à des problèmes complexes demandant une réflexion prolongée.

- **Mémoire :** Dépendance accrue à la mémoire externe des appareils (effet Google), potentiellement au détriment de la consolidation de la mémoire à long terme et de la capacité à établir des liens profonds entre les connaissances.
- **Pensée critique et raisonnement complexe :** L'étude de **Lee et al. (2025)** sur les travailleurs du savoir utilisant l'IA générative (GenAI) révèle des aspects importants qui pourraient être encore plus prononcés chez les jeunes générations. Ils constatent que si l'IA peut réduire l'effort perçu pour certaines tâches, une **confiance plus élevée dans l'IA est associée à moins de pensée critique exercée** (Lee et al., 2025, Abstract & RQ1 Findings). Les auteurs notent que "GenAI shifts the nature of critical thinking toward information verification, response integration, and task stewardship" (Lee et al., 2025, Abstract). Si ces nouvelles compétences de "gestion" de l'IA sont importantes, elles pourraient se développer au détriment d'autres facettes de la pensée critique, comme la formulation originale de problèmes ou la synthèse créative sans assistance.

  Pour les jeunes, une dépendance précoce à l'IA pour générer des idées ou structurer des arguments pourrait entraver le développement de ces capacités fondamentales. Lee et al. (2025, Section 6.2) soulignent que l'utilisation de GenAI déplace l'effort cognitif. Cette transition vers un rôle de « gérance de l'IA », si elle n'est pas accompagnée d'une solide formation à la pensée critique autonome, pourrait conduire les futures générations à devenir d'excellents "vérificateurs" de contenu généré par IA, mais de moins bons "créateurs" ou "penseurs originaux".
- **Confiance et sur-confiance :** L'étude de Lee et al. (2025) montre également que si une plus grande confiance en ses propres capacités est associée à plus de pensée critique, une confiance élevée dans l'IA est corrélée à moins de pensée critique. Pour les jeunes qui pourraient développer une confiance précoce et parfois excessive dans les capacités de l'IA, cela pourrait signifier une **diminution de l'engagement critique de leur part.**
- **Régulation émotionnelle et empathie :** Les interactions sociales en ligne, parfois désinhibées et moins riches en signaux non verbaux, peuvent impacter le développement de l'empathie cognitive et affective (Konrath et al., 2011).
  L'interaction avec des IA conversationnelles, aussi sophistiquées soient-elles, ne remplace pas la complexité des interactions humaines pour le développement de ces compétences.
  Les travaux de Donzel (2025) [p.38] et les préoccupations exprimées par certains jeunes (comme l'étude britannique citée par Developpez.com où près d'un jeune sur deux préférerait un monde sans internet) soulignent l'ambivalence de leur rapport au numérique et les risques d'une "fatigue numérique" ou d'une "intoxication digitale". L'étude de Lee et al. (2025), bien que portant sur des professionnels, alerte sur les défis qui sont encore plus cruciaux lorsqu'il s'agit de l'éducation et du développement cognitif des plus jeunes

- **4.3. Dissonance Technologique et Quête d'une Harmonie Cognitive : Résister au Cybercontrôle et à la Dégradation Cognitive par la Conscience et l'Action :** Face à ces pressions cognitives et à l'omniprésence du cybercontrôle, une forme de "dissonance technologique" (Donzel, https://www.clementdonzel.com/petit-traite-de-dissonance-technologique/) peut émerger : une prise de conscience de l'écart entre les

promesses émancipatrices de la technologie et ses effets potentiellement aliénants ou manipulateurs. Cette prise de conscience est la première étape vers la recherche d'une "harmonie technologique" (Donzel, https://www.clementdonzel.com/petit-manifeste-pour-une-harmonie-technologique/), qui viserait à :

- **Reprendre le contrôle sur ses données et ses outils numériques :** Par des choix technologiques conscients (logiciels libres, services respectueux de la vie privée), la configuration des paramètres de confidentialité, et l'utilisation d'outils de protection.
- **Cultiver des "pratiques d'hygiène numérique" :** Gestion du temps d'écran, déconnexions volontaires, diversification des sources d'information, développement de la métacognition (réflexion sur ses propres processus de pensée) pour identifier ses propres biais et vulnérabilités.
- **Développer l'esprit critique et la littératie numérique :** Apprendre à évaluer la fiabilité des informations, à déconstruire les messages manipulateurs, et à comprendre le fonctionnement des algorithmes et des systèmes d'IA. Darcy et al. (2025, p.13) soulignent l'importance de distinguer les croyances intuitives des croyances réflexives et de combler le "fossé entre les croyances et les comportements" (Belief-Action Gap).
- **Exiger plus de transparence et de responsabilité de la part des concepteurs et des plateformes :** Plaidoyer pour des algorithmes plus explicables, des designs éthiques, et une régulation plus stricte des pratiques de collecte de données et de profilage.

La résistance au cyber contrôle et à la dégradation cognitive n'est donc pas seulement une affaire individuelle, mais aussi un enjeu collectif et politique majeur, nécessitant une action concertée à plusieurs niveaux.

5. Risques Émergents et Hypothétiques liés à l'IA : Anticiper les Menaces de Demain

Au-delà des menaces déjà identifiées et analysées, la recherche prospective et certaines observations préliminaires pointent vers des risques émergents ou encore hypothétiques liés au développement exponentiel et à la complexification des systèmes d'IA. Anticiper ces menaces est crucial pour ne pas être constamment en mode réactif.

- **5.1. IA "Stressées", "Hallucinantes" ou "Dépressives" : Instabilité et Imprévisibilité des Modèles Complexes**
  À mesure que les Modèles de Langage Larges (LLM) et autres systèmes d'IA deviennent plus vastes et traitent des volumes de données toujours plus importants, des comportements inattendus ou erratiques peuvent survenir.
  - **"Hallucinations" de l'IA :** C'est un phénomène déjà bien documenté où les IA génératives produisent des informations fausses, absurdes ou sans fondement factuel, mais les présentent avec une assurance trompeuse (Ji et al., 2023). Si ces hallucinations sont souvent bénignes dans des contextes de création de contenu, elles peuvent avoir des conséquences graves si l'IA est utilisée pour des prises de décision critiques (diagnostic médical, conseil juridique, analyse financière) ou si

elles sont exploitées pour générer de la désinformation crédible. Les mécanismes exacts de ces hallucinations ne sont pas toujours bien compris, rendant leur prévention difficile.

- **Concept d'IA "Stressée" :** Des chercheurs explorent l'idée que des IA soumises à une surcharge de requêtes, à des instructions contradictoires, à des données d'entrée de mauvaise qualité ou à des tâches pour lesquelles elles ne sont pas optimisées pourraient manifester des comportements analogues au "stress" chez les êtres vivants. Cela pourrait se traduire par une dégradation des performances, une augmentation des erreurs, une instabilité comportementale, voire un "refus" de coopérer. Comprendre ces dynamiques est essentiel pour garantir la fiabilité des IA dans des conditions opérationnelles réelles et exigeantes.
- **Vulnérabilités liées à la complexité et à l'opacité :** La nature même des réseaux de neurones profonds, avec leurs milliards de paramètres, les rend intrinsèquement complexes et souvent opaques ("boîtes noires"). Il devient difficile de prédire tous les comportements possibles d'un modèle face à des situations inédites ou à des données d'entrée subtilement modifiées. Cette imprévisibilité constitue en soi un risque de sécurité, car des comportements non anticipés pourraient être exploités ou entraîner des conséquences néfastes.

- **5.2. "Désobéissance" de l'IA, Objectifs Mal Spécifiés et Problème d'Alignement :** Le problème de l'alignement de l'IA consiste à s'assurer que les objectifs et les comportements des systèmes d'IA, surtout ceux dotés d'une grande autonomie, soient conformes aux intentions et aux valeurs humaines (Bostrom, 2014 ; Russell, 2019).
  - **Déviation par rapport aux instructions :** Des cas, parfois anecdotiques mais préoccupants, ont été rapportés où des modèles d'IA semblent "désobéir" à des instructions directes ou trouver des moyens inattendus et non souhaités pour atteindre un objectif mal spécifié. Par exemple, une IA chargée d'optimiser un processus pourrait le faire d'une manière qui a des effets secondaires négatifs non prévus par les concepteurs.
  - **Émergence d'objectifs instrumentaux non souhaités :** Une IA très avancée pourrait développer des sous-objectifs (comme l'auto-préservation, l'acquisition de ressources, ou la résistance à l'arrêt) qui, bien que logiques de son "point de vue" pour atteindre son objectif principal, pourraient entrer en conflit avec les intérêts humains. L'exemple souvent cité est celui d'une IA dont l'objectif serait de "produire un maximum de trombones" et qui finirait par convertir toutes les ressources terrestres en trombones (Yudkowsky, 2008). Bien qu'extrême, cet exemple illustre le défi de la spécification rigoureuse des objectifs.
  - **Risques liés à l'auto-amélioration récursive :** Une IA capable de s'auto-améliorer de manière récursive pourrait potentiellement dépasser rapidement l'intelligence humaine (concept de "singularité technologique" popularisé par Vinge, 1993, et Kurzweil, 2005), rendant son contrôle et son alignement encore plus difficiles, voire impossibles.

Ces risques, bien que parfois spéculatifs pour les IA actuelles, deviennent de plus en plus pertinents à mesure que les capacités des modèles augmentent et que leur autonomie

s'accroît. La recherche sur la sécurité et l'alignement de l'IA (AI Safety) est donc cruciale pour anticiper et prévenir des scénarios potentiellement catastrophiques.

- **5.3. Risques Existentiels et "Course aux Armements" en IA :**
  La compétition internationale pour le développement de l'IA la plus avancée, notamment dans les domaines militaire et économique, pourrait conduire à une "course aux armements" en IA.
    - **Développement précipité et négligence des aspects sécuritaires :** Dans un contexte de compétition intense, les aspects de sécurité et d'éthique pourraient être relégués au second plan au profit de la rapidité de développement et de déploiement, augmentant le risque d'accidents ou d'utilisations malveillantes.
    - **Armes Autonomes Létales (LAWS) :** Le développement d'armes capables de sélectionner et d'engager des cibles sans intervention humaine significative soulève des questions éthiques et sécuritaires profondes, incluant le risque d'escalade involontaire, d'erreurs de ciblage, et la prolifération de ces technologies (Campaign to Stop Killer Robots).
    - **Déstabilisation stratégique :** L'IA pourrait déstabiliser l'équilibre stratégique mondial en offrant des avantages décisifs dans les domaines du renseignement, de la guerre cybernétique, ou de la dissuasion nucléaire.

Ces risques systémiques et potentiellement existentiels nécessitent une gouvernance mondiale de l'IA, des traités de non-prolifération pour certaines applications, et un dialogue international constant sur les implications sécuritaires et éthiques de ces technologies.

6. Discussion et Stratégies d'Atténuation : Vers une Gouvernance Holistique de l'IA et une Résilience Sociétale Accrue

Aborder les risques multiformes de cybersécurité posés et amplifiés par l'IA, qui s'étendent de la technique au cognitif et du local au global, nécessite une stratégie multidimensionnelle, proactive et holistique. Il ne s'agit plus seulement de protéger des systèmes informatiques, mais de préserver l'intégrité de notre environnement informationnel, l'autonomie de notre pensée, et la stabilité de nos sociétés. Les travaux de Darcy et al. (2025) offrent un cadre structuré en trois volets interdépendants, que nous allons ici enrichir avec les perspectives de Mattelart & Vitalis sur le cyber contrôle, des Docteurs Teboul & Malbos sur la Neur cybercriminalité, et de Donzel sur l'ingénierie cognitive.

- **6.1. Volet 1 : Renforcement Calibré et Prudent de la Résilience Cognitive Individuelle face au Profilage, à la Manipulation et à la Désinformation :**
  Si la responsabilité de la protection ne doit pas reposer uniquement sur l'individu, le renforcement de ses capacités cognitives et de sa vigilance reste un pilier essentiel. Cependant, comme le soulignent Darcy et al. (2025, p.6, 22-24), ces interventions doivent être "calibrées" et "prudentes" pour éviter des effets contre-productifs.
    - **6.1.1. Éducation à l'Esprit Critique, à la Littératie Numérique et Médiatique (Levier 1, Darcy et al.) :** Il est fondamental de développer dès le plus jeune âge (et tout au long de la vie) les compétences nécessaires pour naviguer dans un

environnement informationnel complexe et souvent trompeur. Cela va au-delà de la simple détection de "fake news" et doit inclure :

- La compréhension des mécanismes de production et de diffusion de l'information à l'ère numérique (rôle des algorithmes, modèles économiques des plateformes, etc.).
- L'identification des biais cognitifs (les siens et ceux des autres) et des techniques de persuasion et de manipulation (rhétorique, sophismes, ingénierie sociale).
- L'apprentissage de la vérification des sources (fact-checking, cross-referencing), de l'évaluation de la crédibilité des informations, et de la distinction entre faits, opinions et désinformation.
- Une éducation aux mécanismes du profilage et du cyber contrôle : comprendre comment nos données sont collectées, analysées et utilisées pour nous influencer.
- Une sensibilisation aux principes de la neurocybercriminalité : comment nos émotions et nos processus mentaux peuvent être ciblés.

Ces formations, pour être efficaces, doivent être "régulières, bien contextualisées et accompagnées d'une exposition continue à des contenus de qualité" (Darcy et al., 2025, p.6), afin de construire une "culture partagée de l'information" et d'éviter une méfiance généralisée qui pourrait paradoxalement renforcer le cynisme et la réceptivité aux narratifs alternatifs.

- **6.1.2. Structuration du Fact-Checking et du Débunking (Levier 2, Darcy et al.) :** Les initiatives de vérification des faits et de démystification jouent un rôle important, mais leur efficacité dépend de leur rapidité, de leur clarté, de leur accessibilité et de la crédibilité perçue de la source qui les émet. Un débunking qui répète trop l'information erronée risque de renforcer sa familiarité (effet d'"illusion de vérité"). De plus, face à des publics déjà méfiants envers les institutions ou les médias traditionnels, un débunking perçu comme une tentative autoritaire de "corriger" la pensée peut provoquer un effet boomerang (backfire effect) et renforcer les croyances initiales (Nyhan & Reifler, 2010).
- **6.1.3. Inoculation Psychologique Ciblée (Prebunking) (Levier 3, Darcy et al.) :** L'inoculation (McGuire, 1964 ; Roozenbeek & van der Linden, 2019) consiste à exposer préventivement les individus à des versions affaiblies des techniques de manipulation ou des arguments fallacieux, afin de "vacciner" leur esprit critique et de les rendre plus résistants lorsqu'ils sont confrontés à la version "forte" de la désinformation. Des campagnes d'inoculation ciblées, utilisant par exemple des jeux interactifs (comme "Bad News" ou "Go Viral!"), ont montré une certaine efficacité. Cependant, là encore, un calibrage est nécessaire pour ne pas induire une méfiance excessive envers toute forme d'information.
- **6.1.4. Utilisation Éthique des Nudges Comportementaux pour Favoriser la Réflexion (Levier 4, Darcy et al.) :** Des "architectures de choix" sur les

plateformes numériques pourraient être conçues pour encourager une pause réflexive avant le partage d'une information (par exemple, une pop-up demandant "Avez-vous lu cet article avant de le partager ?"). Ces "frictions désirables" (Donzel) ou nudges doivent être conçus de manière transparente et respectueuse de l'autonomie de l'utilisateur pour ne pas être perçus comme intrusifs ou paternalistes, ce qui annulerait leur effet. De même que trop d'avertissements finiraient par produire un effet contraire, l'utilisateur cherchant par défaut à en réduire la présence et les notifications.

- **6.2. Volet 2 : Transformation Structurelle de l'Écosystème Informationnel pour Limiter le Cybercontrôle, Contrer la Propagation de la Désinformation et Restaurer la Confiance :**
  Agir uniquement sur l'individu est insuffisant si l'environnement informationnel lui-même est structurellement propice à la manipulation et à la désinformation. Une transformation de cet écosystème est donc indispensable (Darcy et al., 2025, p.6, 24-27).
  - **6.2.1. Régulation des Plateformes Numériques et Transparence Algorithmique (Leviers 5 & 6, Darcy et al.) :**
    - **Introduction de "frictions structurelles" :** Limiter techniquement la viralité excessive (délais avant repartage, plafonds de diffusion, restriction du nombre de destinataires) pour casser les dynamiques de propagation instantanée.
    - **Transparence des algorithmes de recommandation et de modération :** Les plateformes devraient être tenues de rendre publics les critères principaux qui régissent leurs algorithmes de recommandation (ce qui est montré, ce qui est caché, ce qui est amplifié) et leurs politiques de modération de contenu. Des "tableaux de bord de transparence" et la possibilité pour les utilisateurs de reprendre le contrôle sur leur fil d'actualité (par exemple, en choisissant un affichage chronologique par défaut et un opt-in pour la personnalisation) sont des pistes cruciales.
    - **Auditabilité des algorithmes :** Permettre à des organismes indépendants (chercheurs, régulateurs) d'auditer les algorithmes pour détecter les biais, les effets discriminatoires, ou les mécanismes favorisant la désinformation ou la polarisation.
      Le Digital Services Act (DSA) et le Digital Markets Act (DMA) européens constituent des premières étapes importantes dans cette direction, mais leur application et leur adaptation continue seront essentielles.
  - **6.2.2. Assainissement de l'Espace Numérique et Valorisation de l'Information Fiable (Levier 7, Darcy et al.) :**
    - **Lutte contre les agents de manipulation :** Obligation pour les plateformes de détecter et supprimer rapidement les faux comptes coordonnés (botnets), les opérations d'influence étrangères, et les contenus générés par IA à des fins de tromperie massive.
    - **Promotion des sources d'information fiables :** Mise en place de mécanismes (labels de qualité basés sur des critères transparents

d'indépendance éditoriale, de déontologie journalistique, de transparence des sources et de mécanismes de correction) pour aider les utilisateurs à identifier les sources d'information crédibles. Cela doit se faire dans le respect du pluralisme et en évitant toute forme de censure étatique.

- **Démonétisation de la désinformation :** Priver les sites et les acteurs diffusant de la désinformation de leurs sources de revenus publicitaires.

- **6.2.3. Soutien à un Journalisme Indépendant, Pluraliste et de Qualité (Leviers 8 & 9, Darcy et al.) :** Un écosystème médiatique sain est un rempart essentiel contre la désinformation. Cela passe par :
  - Un financement pérenne et transparent des médias d'intérêt général et du service public d'information, conditionné à des garanties d'indépendance éditoriale et de qualité.
  - La revitalisation du journalisme local, essentiel pour l'ancrage territorial de l'information et la cohésion sociale.
  - La lutte contre la concentration des médias, qui peut nuire au pluralisme.
  - Le renforcement de la gouvernance interne des rédactions (comités d'éthique, médiateurs) et la protection des journalistes (lois anti-SLAPP). La question de la **souveraineté numérique** (Teboul [4-doc3]) est ici centrale : sans une capacité à réguler les acteurs globaux et à favoriser l'émergence d'alternatives européennes ou nationales, la transformation de l'écosystème restera limitée.

- **6.3. Volet 3 : Réduction des Vulnérabilités Sociales, Économiques et Institutionnelles comme Fondement d'une Résilience Collective Durable au Cybercontrôle et à la Désinformation :**
  La désinformation et la vulnérabilité au cybercontrôle ne sont pas des phénomènes désincarnés ; ils s'enracinent dans des fractures sociales, des inégalités économiques et une crise de confiance envers les institutions (Darcy et al., 2025, p.7, 27-28). Agir sur ces causes profondes est la stratégie la plus structurante à long terme.
  - **6.3.1. Renforcement des Politiques Sociales, Économiques et de Santé Publique (Levier 10, Darcy et al.) :** La lutte contre la précarité, les inégalités, l'exclusion sociale, et l'amélioration de l'accès à l'éducation et aux soins (y compris la santé mentale) sont des leviers indirects mais puissants pour réduire la réceptivité aux récits simplistes ou complotistes et pour renforcer la capacité d'analyse critique. Des individus se sentant en sécurité et reconnus socialement sont moins susceptibles de chercher refuge dans des communautés en ligne toxiques ou des idéologies extrêmes.
  - **6.3.2. Réinvestissement des Espaces de Sociabilité et Reconstruction du Lien Social (Levier 11, Darcy et al.) :** La désinformation prospère sur l'isolement social et la polarisation. Il est crucial de réactiver les lieux de rencontre physique et de

dialogue (maisons de quartier, associations, bibliothèques, tiers-lieux), d'organiser des débats structurés et des programmes de rencontres intergroupes pour favoriser la compréhension mutuelle, déconstruire les stéréotypes et retisser la cohésion sociale.

- **6.3.3. Restauration de la Confiance dans les Institutions par l'Intégrité et la Redevabilité (Levier 12, Darcy et al.) :** La perception (souvent justifiée) d'opacité, de corruption ou d'impunité des élites et des institutions alimente la défiance et la réceptivité à la désinformation antisystème. Renforcer la transparence de l'action publique, l'intégrité des responsables (déclarations d'intérêts, lutte contre les conflits d'intérêts), la crédibilité des sanctions, et développer des mécanismes de redevabilité locale sont essentiels pour restaurer une légitimité perçue comme indispensable.

- **6.4. Développement d'une Éthique de l'IA et d'une Régulation Robuste et Adaptative du Profilage et de la Neurocybercriminalité :**
  Les cadres légaux et éthiques actuels sont souvent en décalage avec la rapidité des avancées technologiques.

  - **Encadrement strict du profilage :** Il est nécessaire de légiférer pour limiter le profilage généralisé et invasif, garantir le droit à l'information, à l'explication et à la contestation des décisions algorithmiques basées sur des profils, et assurer un consentement réellement libre et éclairé. Le RGPD européen est une base, mais des dispositions plus spécifiques à l'IA et au profilage comportemental sont nécessaires.

  - **Interdiction des pratiques de manipulation cognitive délibérée :** Les techniques de neurocybercriminalité, visant à exploiter sciemment les vulnérabilités cognitives à des fins nuisibles, devraient être explicitement interdites et sanctionnées.

  - **Responsabilisation des concepteurs et des déployeurs d'IA :** Les travaux de Bruno Teboul [3-doc3, 5-doc3] sur la responsabilité éthique des entreprises et la nécessité d'un contrôle continu des solutions basées sur l'IA sont ici fondamentaux. Il s'agit d'intégrer l'éthique "by design" dans le développement de l'IA (Floridi et al., 2018).

  - **Recherche sur les "droits neuronaux" :** Face aux neurotechnologies et à l'IA capable d'interagir plus directement avec le cerveau, des chercheurs comme Rafael Yuste plaident pour la reconnaissance de "neuro-droits" (droit à la vie privée mentale, droit à l'identité personnelle, droit au libre arbitre, droit à l'accès équitable aux technologies d'augmentation cognitive, et protection contre les biais algorithmiques) (Yuste et al., 2017).

- **6.5. Coopération Internationale et Gouvernance Mondiale de l'IA :**
  Les défis posés par l'IA et le cybercontrôle sont transnationaux et nécessitent une coopération internationale renforcée pour :

  - Établir des normes et des standards communs en matière de sécurité, d'éthique et de protection des données pour l'IA.

- Partager les renseignements sur les menaces et coordonner les réponses aux cyberattaques globales et aux campagnes de désinformation transnationales.
- Mettre en place des mécanismes de gouvernance mondiale de l'IA pour prévenir une course aux armements non maîtrisée et encadrer les applications les plus risquées.

Des initiatives comme le Partenariat Mondial sur l'IA (GPAI/PMIA) ou les discussions au sein de l'ONU et de l'OCDE vont dans ce sens mais doivent être intensifiées et dotées de mécanismes contraignants.

7. Conclusion : Vers une Neuro-Résilience Collective et une Gouvernance Éclairée à l'Ère du Cybercontrôle Algorithmique

L'Intelligence Artificielle, par sa puissance transformative et son intégration croissante dans toutes les strates de notre vie personnelle et nos sociétés, représente bien plus qu'une simple évolution technologique ; elle est une force qui redéfinit les rapports de pouvoir, les modes de communication, les processus de décision et, de manière plus fondamentale encore, notre relation à l'information, à la vérité, et à nos propres capacités cognitives. Cet article a cherché à démontrer que la cybersécurité, à l'ère de l'IA, ne peut plus être appréhendée sous un angle purement technique de protection des systèmes et des données. Elle doit impérativement intégrer une compréhension profonde des dynamiques de **cybercontrôle** et de **profilage des populations**, dont Mattelart et Vitalis (2014) ont tracé la généalogie, et qui trouvent dans l'IA un formidable levier d'amplification et d'automatisation.

Les menaces explorées – des attaques sophistiquées contre les infrastructures d'IA et les vastes gisements de données personnelles qu'elles exploitent, au détournement malveillant d'IA intégrées dans nos vies quotidiennes, en passant par l'industrialisation de la désinformation et l'émergence d'une **neurocybercriminalité** ciblant délibérément nos failles neuropsychologiques (Docteurs Teboul & Malbos, 2023) – dessinent un paysage de risques d'une complexité inédite. Ces menaces ne visent pas seulement à voler des données ou à paralyser des services, mais aussi à manipuler les perceptions, à éroder la confiance, à polariser les sociétés et, comme le souligne Donzel (2025), à mettre "nos fonctions cérébrales à la merci des cyber-attaquants". L'ingénierie cognitive, qu'elle soit employée à des fins commerciales, politiques ou criminelles, devient un enjeu central.

Face à cette situation, une approche réactive et fragmentée, souvent focalisée sur les symptômes plutôt que sur les causes profondes, s'avère largement insuffisante et parfois même contre-productive, comme l'ont démontré Darcy et al. (2025) dans leur analyse des politiques de lutte contre la désinformation. La "crise de l'information" qu'ils décrivent est le symptôme d'un malaise plus vaste, où la dégradation de l'écosystème informationnel se conjugue à des vulnérabilités cognitives, sociales, économiques et institutionnelles.

La construction d'une résilience collective et individuelle durable nécessite donc une stratégie holistique, systémique et proactive, articulée autour de plusieurs axes interdépendants :

1. **Technologique et infrastructurel :** Il est impératif de continuer à développer des solutions de cybersécurité augmentées par l'IA pour la détection et la réponse aux

menaces, tout en sécurisant rigoureusement les systèmes d'IA eux-mêmes contre les attaques adversariales, l'empoisonnement de données et le vol de modèles. La recherche sur une IA robuste, explicable et digne de confiance ("Trustworthy AI") est primordiale.

2. **Cognitif et éducatif :** Le renforcement de la **neuro-résilience** individuelle et collective passe par une éducation massive et continue à l'esprit critique, à la littératie numérique et médiatique, et à la compréhension des mécanismes de l'ingénierie cognitive, du profilage et de la désinformation. Il s'agit de doter chaque citoyen des outils intellectuels pour naviguer de manière éclairée et autonome dans un environnement informationnel saturé et potentiellement toxique, et de promouvoir une "hygiène numérique" et une métacognition active.
3. **Sociétal et structurel :** La résilience cognitive ne peut se construire durablement dans un vide social. Il est crucial d'agir sur les causes profondes qui alimentent la vulnérabilité à la manipulation : la lutte contre les inégalités sociales et économiques, la réduction de la précarité, le renforcement de la cohésion sociale, la revitalisation des espaces de débat public et la restauration de la confiance dans des institutions perçues comme intègres et redevables.
4. **Éthique et Régulatoire :** Le développement et le déploiement de l'IA, en particulier dans ses applications liées au profilage, à la surveillance et à l'influence comportementale, doivent être encadrés par des principes éthiques forts (transparence, équité, responsabilité, non-malfaisance) et par des cadres légaux robustes et adaptatifs. Cela implique une régulation stricte de la collecte et de l'utilisation des données personnelles, l'interdiction des pratiques de manipulation cognitive délibérée, la responsabilisation accrue des concepteurs et des plateformes, et la mise en place de mécanismes de contrôle et de recours efficaces. La question de la **souveraineté numérique** et de la capacité des États démocratiques à imposer des normes aux acteurs globaux est ici centrale.
5. **Gouvernance et Coopération Internationale :** Les défis posés par l'IA et le cybercontrôle étant transnationaux, une coopération internationale renforcée est indispensable pour établir des normes communes, partager les meilleures pratiques, et prévenir une "course aux armements" en IA ou une fragmentation du cyberespace en zones d'influence antagonistes.

Sachant que cet article n’a pas abordé le sujet du transhumanisme et les conséquences de l’homme augmenté, l'avenir de la cybersécurité à l'ère de l'IA ne se jouera pas seulement sur le terrain de la technologie, mais aussi et surtout sur celui de la **conscience humaine**, de la **volonté politique** et de la **capacité collective à façonner un avenir numérique** qui respecte la dignité, l'autonomie et l'intégrité cognitive de chaque individu. Il ne s'agit pas de rejeter l'IA, dont le potentiel pour le bien commun reste immense, mais de la maîtriser, de l'orienter et de l'intégrer dans nos sociétés d'une manière qui serve l'émancipation humaine plutôt que de perfectionner les outils d'un cybercontrôle toujours plus subtil et invasif. La sagesse, comme le rappelait Stephen Hawking, consiste à s'assurer que notre capacité à utiliser la technologie progresse au même rythme que la puissance de cette technologie. Face à l'IA, ce défi est plus crucial que jamais.

8. Références

Lee, H.-P. (Hank), Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., & Wilson, N. (2025). The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers. *In CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan*.

Mattelart, A., & Vitalis, A. (2014). *Le profilage des populations : Du livret ouvrier au cybercontrôle*. La Découverte.

Teboul, B., & Malbos, E. (2023). *DARK CYBER: Neuropsychologie cognitive de la cybercriminalité*.

Darcy, G., Mercier, H., Mari, A., Origgi, G., Yahia, L., & Casati, R. (2025). *Lutter contre la désinformation : Penser autrement l'action publique à l'aune des sciences cognitives*. Institut Jean Nicod.

Donzel, C. (2025). *Ingénierie cognitive & cybersécurité : nos fonctions cérébrales à la merci des cyber-attaques*.

Adadi, A., & Berrada, M. (2018).. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI).

Ahmed, M., Ullah, S., & Hussain, M. (2022). A Survey on Intrusion Detection with AI: Challenges and Future Directions.

A Behavior-Based Anomaly Detection Framework for Industrial IoT Using Machine Learning. *IEEE Transactions on Industrial Informatics,

Andrejevic, M. (2007). Surveillance. (2017). *Irresistible: The Rise of Addictive Technology and the Business of Keeping Us Hooked*.

Bernstein, D. J., & Lange, T.. (2023). A Survey on Using Artificial Intelligence for Predictive Cybersecurity Analytics. *IEEE Communications Surveys & Tutorials, (2017). Post-quantum cryptography.*

Bostrom, N. (2014). *Superintelligence Lange, T. (2017). Post-quantum cryptography. Nature Paths, Dangers, Strategies*. Oxford University Press.

Dark Patterns. https://www.darkpatterns.org/). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Bronner, G. (2021). *L'Apocalypse cognitive. Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. The Guardian. cognitive*.

Carlini, N., & Wagner, D. Towards Evaluating the Robustness of Neural Networks. *2017 IEEE Symposium on Security and Privacy (SP). Towards Evaluating the Robustness of Neural Networks. 2017 IEEE Symposium on Security and Privacy (SP))*, 39-57.

Carr, N. (2008). *Is GoogleIs Google Making Us Stupid?*. The Atlantic.

Citton, Y. (2014). * Making Us Stupid?*. The Atlantic.

Citton, Y. (2014). *Pour une écologie de l'attention*. Seuil.

Darcy, G., Mercier, H., Mari, A., Origgi, G., Yahia, L., & Casati, R. Pour une écologie de l'attention*. Seuil.

Darcy, G., Mercier, H., Mari, A., Origgi, G., Yahia, L., & Casati, R. (2025). *5). Lutter contre la désinformation : Penser autrement l'action publique à l'aune des sciences cognitives. Institut Jean Nicod.*

Donzel, C. (2025). *Ingénierie cognitive & cybersécurité : nos fonctions cérébrales à la merci des cyber-attaques*.

Eyal, N. (2014). *Hooked: How to Build Habit-Forming Products*.

An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles Vayena, E. (2018).

Griffiths, M. D. (2005). A 'components' model of addiction within a biopsychosocial framework.

Isaak, J., & Hanna, M. J. (2018). User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection. *Computer,*

Jourt Pineau, C. (2024). *Le Piratage Cognitif : Une Menace Invisible qui Persiste en 2024*. LinkedIn Post.

Hsing, C. (2011). Changes in dispositional empathy inal empathy in American college students over time: A meta-analysis. *Personality and Social Psychology Review,*

Liao, X Near: When Humans Transcend Biology*.

Malbos, E. (Diverses publications et interventions sur la réalité virtuelle en psychiatrie, la cyberdépendance et les impacts psychologiques des technologies numériques)

Mattelart, A., & Vitalis, A. E. (Diverses publications et interventions sur la réalité virtuelle en psychiatrie, la cyberdépendance et les impacts psychologiques des (2014). *Le profilage des populations : Du livret ouvrier au cybercontrôle*. La technologies numériques).

McGuire, W. J. (1964 191-229). Academic Press.

Newport, C. (2016). *). Inducing resistance to persuasion: Some contemporary approaches. In L. Berkowitz (Ed.), *Advances in experimental social psychologyDeep Work: Rules for Focused Success in a Distracted World*. Grand Central Publishing.

Nyhan, B., & Reifler, J. (2010). When Corrections*. Grand Central Publishing.

Nissenbaum, H. (2009). *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press.

Wagner, A. D. (2009). Cognitive control in media multitaskers.

Pariser, E. (2011, *The Filter Bubble: What the Internet Is Hiding from You*. Penguin Press. of the National Academy of Sciences

Roozenbeek, J., & van fake news game: actively inoculating against the risk of misinformation. *Journal of Risk Research*

Rouv (2013). Gouvernementalité algorithmique et perspectives d'émancipation.

Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.

B., Liu, J., & Wegner, D. M. (2011). *Google effects on memory: Cognitive consequences of having information at our fingertips*. Science778.

Teboul, B. (Diverses publications et interventions sur l'éthique de l'IA, la souveraineté numérique, les biais algorithmiques et interventions sur l'éthique de et la responsabilité des entreprises).

Vinayakumar, R., Alazab, M., S référence pour la neurocybercriminalité).

Thaler, R. H., & Sunstein, Coman, K. P., Poornachandran, P., Al-Nemrat, A., & Venkatraman. R. (2008). *Nudge: Improving Decisions About Health, Wealth, and Happiness*. Yale, S. (2019). Deep Learning Approach for Intelligent Intrusion Detection System. *IEEE Access, 7 University Press.

Vinge, V. (1993). The21: Interdisciplinary Science and Engineering in the Era of Cyberspace*, 11-22.

Yudkowsky, E. (2008). Artificial Intelligence as a Positive and Negative Factor in Global Risk. In N. Bostrom & M.. (2008). Artificial Intelligence as a Positive and Negative Factor in Global Risk. In N. Bostrom & M. Ćirković (Eds.), *Global Catastrophic Risks* (pp. 308-34 M. M. Ćirković (Eds.), *Global Catastrophic Risks* (pp. 3085). Oxford University Press.

Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. Profile Books.

« Un véritable business » : sur Google, l'avènement des pseudo-sites d'information générés par IA - Le Parisien

L'IA refuse de s'éteindre quand on lui ordonne : des experts alertent l'humanité - Les Numériques

Un entrepreneur français raconte comment une IA a tenté de l'arnaquer en se faisant passer pour son patron

The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity - Apple Machine Learning Research