

# LA SOUVERAINETÉ COGNITIVE À L'HEURE DE L'IA ET DES GRANDS MODÈLES DE LANGAGE (LLM)

---

**M. GUIDERE.** Professeur des Universités (Paris 8). Directeur de recherches à l'INSERM

*Institut National de la Santé et de la Recherche Médicale, France.*

[mathieu.guidere@inserm.fr](mailto:mathieu.guidere@inserm.fr)

---

**Résumé :** Cet article examine la notion de souveraineté cognitive à l'ère de l'intelligence artificielle générative, en se concentrant sur l'impact des grands modèles de langage sur les processus mentaux individuels et collectifs. En mobilisant des apports issus des sciences cognitives, des neurosciences, des sciences du langage et des sciences sociales, il met en évidence les transformations profondes induites par ces technologies dans les modalités de raisonnement, de décision, de créativité et de formation des opinions. L'analyse souligne les risques associés à la délégation cognitive, notamment la dépendance fonctionnelle, l'atrophie de l'engagement neuronal, la dette cognitive et la perte progressive d'autonomie décisionnelle.

L'article explore également les dynamiques d'influence, de manipulation algorithmique et de guerre cognitive, en montrant comment les LLM peuvent être intégrés à des stratégies de polarisation, de désinformation et de déstabilisation démocratique. Enfin, il propose un cadre conceptuel pour penser la souveraineté cognitive comme une stratégie nationale, articulant protection des neurodroits et renforcement des capacités critiques par l'éducation. L'ensemble vise à contribuer à une réflexion critique sur les conditions de préservation de l'autonomie mentale dans un environnement technologique en mutation rapide.

**Mots clés :** Souveraineté cognitive, intelligence artificielle générative, grands modèles de langage, dépendance cognitive, influence algorithmique, neurodroits, démocratie, guerre cognitive.

---

**Abstract:** This article examines the concept of cognitive sovereignty in the era of generative artificial intelligence, with a particular focus on the impact of large language models on individual and collective mental processes. Drawing on contributions from cognitive science, neuroscience, language sciences, and the social sciences, it highlights the profound transformations induced by these technologies in modes of reasoning, decision making, creativity, and opinion formation. The analysis emphasizes the risks associated with cognitive delegation, including functional dependency, reduced neural engagement, cognitive debt, and the gradual erosion of decision making autonomy.

The article also explores dynamics of influence, algorithmic manipulation, and cognitive warfare, showing how large language models can be integrated into strategies of polarization, disinformation, and democratic destabilization. Finally, it proposes a conceptual and normative framework for understanding cognitive sovereignty as a national strategy, combining legal regulation, protection of neurorights, and the strengthening of critical capacities through education. The overall objective is to contribute to a critical reflection on the conditions required to preserve mental autonomy within a rapidly evolving technological environment.

**Keywords:** cognitive sovereignty, generative artificial intelligence, large language models, cognitive dependency, algorithmic influence, neurorights, democracy, cognitive war.

---

## Introduction

La souveraineté cognitive est une notion multidimensionnelle située à l'intersection de la géopolitique, des sciences cognitives et de l'intelligence artificielle. Elle désigne la capacité d'un individu, d'une organisation ou d'une collectivité à conserver la maîtrise de ses processus mentaux, de ses cadres interprétatifs et de ses productions symboliques face à des dispositifs techniques de plus en plus autonomes.

Dans un contexte de transformation numérique accélérée, cette notion prend une importance particulière en raison de l'intégration croissante de systèmes algorithmiques dans les activités de compréhension, de décision et de communication. La souveraineté cognitive ne se limite pas à la protection des données ou à l'indépendance technologique, mais engage des enjeux plus profonds relatifs à l'autonomie épistémique, à la formation des croyances et à la structuration de l'attention.

L'émergence de l'intelligence artificielle générative, et plus spécifiquement des grands modèles de langage, introduit une rupture dans les modalités traditionnelles de production et de circulation du savoir. Ces systèmes, entraînés sur des volumes massifs de textes, sont désormais capables de générer des énoncés cohérents, d'assister le raisonnement humain et d'influencer les pratiques cognitives quotidiennes. Cette évolution soulève des interrogations majeures quant à la délégation de fonctions cognitives autrefois réservées aux sujets humains et quant aux conditions dans lesquelles cette délégation peut affecter la capacité des individus et des sociétés à penser par eux-mêmes.

La souveraineté cognitive apparaît ainsi comme un cadre conceptuel permettant d'analyser les rapports des forces cognitives induits par l'intelligence artificielle. Elle invite à examiner les asymétries entre les concepteurs des systèmes, les détenteurs des infrastructures computationnelles et les utilisateurs finaux, ainsi que les effets de ces asymétries sur la formation des représentations mentales et des normes discursives. Cela est d'autant plus important que le pouvoir de l'intelligence est aujourd'hui détenu par un petit nombre d'acteurs privés, ayant un agenda politique et culturel avéré.

Dans cette perspective, la compréhension de la révolution cognitive portée par les grands modèles de langage constitue un préalable nécessaire à toute réflexion sur les conditions d'une gouvernance cognitive équilibrée à l'ère de l'IA.

## La révolution cognitive des grands modèles de langage

Les grands modèles de langage marquent un tournant dans l'histoire des technologies cognitives. Contrairement aux systèmes symboliques classiques, ces modèles reposent sur des architectures neuronales capables d'extraire des régularités statistiques complexes à partir de vastes corpus linguistiques. Leur fonctionnement ne se fonde pas sur des règles explicites, mais sur l'apprentissage de distributions probabilistes permettant de prédire et de générer des séquences linguistiques adaptées au contexte. Cette approche confère aux LLM une flexibilité inédite et une capacité d'adaptation qui les rapproche, dans certaines tâches, de performances associées à la cognition humaine.

Cette révolution qui touche l'infrastructure de la pensée s'accompagne d'un changement qualitatif dans l'usage du langage comme interface cognitive. Les LLM ne se contentent pas de traiter l'information, ils participent à la co-construction du sens en interaction avec les utilisateurs. En produisant des reformulations, des synthèses ou des hypothèses interprétatives, ils interviennent directement dans les processus de compréhension et de raisonnement. Le langage devient alors un espace partagé entre agents humains et artificiels, où se négocient des significations et des orientations cognitives. Cette hybridation modifie les dynamiques de l'apprentissage, de la créativité et de la prise de décision.

La révolution cognitive induite par les grands modèles de langage se caractérise également par une externalisation accrue de fonctions mentales telles que la mémoire de travail, l'accès aux connaissances et la planification discursive. Cette externalisation peut soutenir les capacités humaines en réduisant certaines charges cognitives, mais elle peut aussi entraîner une dépendance structurelle aux systèmes techniques qui médiatisent l'accès au savoir. À mesure que les LLM s'intègrent dans les environnements professionnels, éducatifs et institutionnels, ils contribuent à redéfinir les normes de compétence cognitive et les attentes en matière de performance intellectuelle.

Enfin, cette révolution s'inscrit dans un contexte globalisé où les grands modèles de langage sont majoritairement développés et contrôlés par un nombre restreint d'acteurs industriels et étatiques. Les choix opérés lors de l'entraînement de ces modèles, qu'il s'agisse des corpus utilisés, des langues privilégiées ou des objectifs d'optimisation, ont des

répercussions directes sur les formes de pensée qu'ils favorisent. La révolution cognitive fondée sur les LLM ne peut donc être comprise indépendamment des enjeux de pouvoir et de gouvernance qui traversent l'écosystème de l'intelligence artificielle, et qui conditionnent l'exercice effectif de la souveraineté cognitive.

## Grands modèles de langage et souveraineté cognitive

La souveraineté cognitive peut être comprise comme l'aptitude d'un sujet individuel ou collectif à maintenir la maîtrise de ses activités mentales, de ses choix intentionnels et de ses états affectifs face à des dispositifs susceptibles d'en orienter le fonctionnement. Elle renvoie à une exigence d'autodétermination psychique, dans laquelle les processus de pensée et de décision demeurent gouvernés par des finalités internes plutôt que par des contraintes imposées de l'extérieur. Cette problématique s'est progressivement imposée dans le débat scientifique et politique à mesure que les technologies numériques ont acquis la capacité d'interagir de manière fine avec l'attention, la mémoire et les mécanismes motivationnels.

L'essor des grands modèles de langage confère une actualité particulière à cette notion, dans la mesure où ces systèmes interviennent directement dans les espaces symboliques où se construisent les représentations et les jugements. En médiatisant l'accès à l'information, en proposant des formulations prêtes à l'emploi et en orientant implicitement les cadres interprétatifs, ils participent à la structuration de l'activité cognitive quotidienne. Leur influence ne relève pas nécessairement d'une contrainte explicite, mais s'exerce souvent de manière diffuse, à travers des suggestions, des priorisations et des régularités discursives qui façonnent les trajectoires de pensée.

À l'échelle individuelle, la question de la souveraineté cognitive concerne la faculté de conserver une initiative réflexive sur ses propres raisonnements. L'assistance apportée par les modèles de langage peut soutenir l'élaboration intellectuelle, mais elle comporte également le risque d'un affaiblissement progressif de certaines compétences métacognitives si l'utilisateur délègue de façon systématique l'analyse, la formulation ou l'évaluation critique. La frontière entre soutien cognitif et substitution fonctionnelle devient alors un enjeu central, notamment lorsque les mécanismes d'influence demeurent opaques ou insuffisamment compris par les usagers.

À l'échelle collective, les grands modèles de langage

soulèvent des enjeux relatifs à la cohérence mentale des groupes sociaux et des communautés politiques. En amplifiant certains récits, en standardisant des formes d'expression ou en facilitant la diffusion de contenus orientés, ils peuvent contribuer à des dynamiques d'homogénéisation cognitive ou, au contraire, à la polarisation des représentations. La capacité d'une société à préserver des espaces de délibération autonome et de pluralité interprétative dépend alors des conditions dans lesquelles ces systèmes sont déployés, régulés et intégrés aux infrastructures informationnelles. Dans ce cadre, la souveraineté cognitive apparaît comme un principe structurant pour penser les rapports entre intelligence artificielle, liberté mentale et autonomie collective.

## État des lieux

L'analyse de la souveraineté cognitive ne saurait se réduire à la seule question de la territorialité des infrastructures ou de la localisation des centres de données. Elle engage plus fondamentalement notre aptitude à élaborer une pensée, à poser un diagnostic et à exercer un arbitrage sans que les processus mentaux ne soient infléchis par des systèmes de valeurs, des biais axiomatiques ou des matrices linguistiques exogènes. Dans cette perspective, l'examen des grands modèles de langage actuels révèle des disparités importantes dans leur rapport à l'autonomie de l'utilisateur.

En effet, la souveraineté cognitive, entendue comme l'aboutissement de l'indépendance numérique, repose sur un triptyque conceptuel dont le premier pilier est l'autonomie de jugement. Celle-ci exige le maintien d'une capacité critique face à des architectures algorithmiques dont les règles de fonctionnement et les processus inférentiels demeurent souvent opaques. S'y ajoute une exigence de préservation culturelle qui vise à prémunir le sujet contre l'imposition de visions du monde, qu'elles soient morales, politiques ou historiques, propres à la zone géographique de conception, à l'instar des paradigmes normatifs issus de la Silicon Valley. Enfin, l'intégrité linguistique garantit la possibilité de déployer une pensée dans sa langue vernaculaire pour échapper à la standardisation induite par la traduction automatique des schèmes de pensée anglo-saxons.

L'examen des systèmes issus des écosystèmes nord-américains, tels qu'OpenAI, Google ou Meta, met en évidence un risque d'homogénéisation cognitive. Entraînés sur des corpus prépondérants en langue anglaise, ces modèles tendent à universaliser des

protocoles de sécurité et des cadres éthiques strictement ancrés dans la culture étasunienne. Il en résulte une forme de dépendance où l'utilisateur, par un phénomène d'ajustement mimétique, finit par formater son propre raisonnement pour l'aligner sur les réponses jugées conformes par l'interface.

À l'inverse, l'alternative française et européenne, représentée par des acteurs comme Mistral AI ou Aleph Alpha, se distingue par une attention portée à la nuance et à l'héritage logique continental. En privilégiant une certaine neutralité et une compréhension native des structures linguistiques, ces modèles limitent l'effet de calque culturel. De plus, le recours au modèle de poids ouverts favorise une réappropriation souveraine en permettant aux organisations d'ancrer l'IA dans des corpus de connaissances locaux, garantissant ainsi une adéquation entre le fonctionnement du modèle et les normes juridiques ou intellectuelles du territoire.

Parallèlement, le bloc asiatique, emmené par des systèmes comme DeepSeek ou Alibaba, propose une alternative radicale caractérisée par une optimisation des capacités logico-mathématiques. Toutefois, si ces architectures rompent avec l'hégémonie occidentale, elles introduisent une autre forme d'aliénation. La souveraineté cognitive y est alors contrainte par des cadres de censure étatiques et des finalités stratégiques qui orientent les trajectoires discursives des modèles.

Ainsi, les différentes architectures révèlent des disparités importantes. Alors que les modèles américains présentent une faible souveraineté culturelle et un risque élevé de dépendance fonctionnelle, les architectures européennes offrent une relative souveraineté de la donnée et une autonomie de décision renforcée par leur auditabilité. Les modèles du Moyen-Orient, à l'instar de Falcon, parviennent à un équilibre par leur ancrage dans les valeurs du Sud Global, tandis que les modèles chinois, malgré une haute spécificité culturelle, demeurent limités par une censure structurelle qui entrave la pleine autonomie de décision de l'utilisateur.

## Risques et périls

Le déploiement massif des grands modèles de langage introduit une série de tensions qui affectent directement l'exercice de l'autonomie mentale dans les environnements numériques contemporains. En tant que technologies cognitives à large spectre, ces systèmes ne se limitent pas à fournir des outils d'assistance, mais interagissent avec des dimensions profondes de l'activité humaine, touchant à la régulation attentionnelle, à la dynamique

motivationnelle et aux mécanismes d'élaboration symbolique. Les enjeux associés dépassent ainsi le cadre technique pour engager des conséquences d'ordre neurocognitif, psychologique, culturel et politique.

Les capacités génératives des modèles de langage leur permettent d'intervenir dans la formation des représentations, l'orientation du raisonnement et la cristallisation des jugements collectifs. En proposant des réponses structurées, des interprétations plausibles et des narrations cohérentes, ils peuvent infléchir les trajectoires de pensée sans recourir à des formes explicites de contrainte. Cette influence indirecte, souvent intégrée dans des usages ordinaires, affecte la créativité, la pensée critique et la construction des opinions publiques, en favorisant certaines associations conceptuelles au détriment d'autres. À ce niveau, le risque réside moins dans une manipulation manifeste que dans une modulation continue et peu perceptible des cadres cognitifs.

À ces effets s'ajoute un problème structurel lié à l'opacité des systèmes intelligents. Les architectures propriétaires, protégées par des logiques industrielles et commerciales, rendent difficile l'analyse des mécanismes internes de génération des contenus. Cette absence de visibilité limite la capacité des utilisateurs, des chercheurs et des institutions à évaluer les biais, les orientations implicites ou les effets cumulatifs de ces modèles. La concentration du savoir algorithmique entre les mains de quelques acteurs contribue ainsi à un déséquilibre cognitif, dans lequel le pouvoir de structurer l'information échappe en grande partie au contrôle démocratique.

Dans les contextes organisationnels et sociaux, l'intégration des grands modèles de langage transforme en profondeur les pratiques communicationnelles et le rapport au savoir. L'automatisation partielle de l'expression écrite et de l'analyse informationnelle peut conduire à une homogénéisation des formes discursives et à une réduction de la variété des points de vue. Cette standardisation cognitive, déjà observée dans certains environnements professionnels, s'accompagne de risques psychosociaux liés à la perte de sens, à la dévalorisation de l'expertise humaine et à l'érosion de la singularité intellectuelle. Ainsi, l'appauvrissement progressif de la diversité des idées constitue l'un des périls les plus insidieux de l'adoption non critique des modèles de langage, en ce qu'il affecte durablement la capacité des collectifs à penser autrement.

## La question des biais cognitifs

Les biais cognitifs, décrits de longue date par les sciences comportementales, constituent des raccourcis mentaux permettant de prendre des décisions rapides en situation d'incertitude. Toutefois, lorsqu'ils sont exploités de manière intentionnelle par des dispositifs algorithmiques adaptatifs, ils deviennent des leviers d'influence susceptibles d'altérer l'autonomie décisionnelle et le discernement critique.

L'évolution récente des technologies d'intelligence artificielle ouvre la voie à des formes d'influence encore plus fines, fondées sur l'analyse directe des réponses neurophysiologiques. Certaines entreprises expérimentent des dispositifs capables de mesurer des indicateurs cérébraux ou biométriques, tels que les variations de l'attention ou de l'émotion, afin d'ajuster en temps réel le contenu publicitaire présenté à une personne. Dans ce cadre, la stimulation n'est plus uniquement calibrée à partir de données comportementales passées, mais s'appuie sur l'état cognitif instantané du sujet. Une telle approche tend à court-circuiter les mécanismes réflexifs en adaptant le message au moment précis où la réceptivité est maximale.

L'enjeu posé par ces pratiques ne réside pas uniquement dans leur efficacité commerciale, mais dans leur capacité à transformer les biais cognitifs en instruments systématiques de modulation mentale. Lorsque ces procédés sont intégrés à des systèmes automatisés capables d'apprentissage et d'optimisation continue, ils instaurent une asymétrie profonde entre les capacités d'adaptation de la machine et les ressources métacognitives de l'individu. La question des biais cognitifs devient alors indissociable de celle de la souveraineté cognitive, dans la mesure où l'exploitation ciblée de ces mécanismes fragilise la possibilité même d'un choix éclairé et autonome.

L'examen des biais cognitifs inhérents aux architectures américaines et européennes des modèles de langage révèle des divergences structurelles majeures dans ce que l'on pourrait qualifier de logiciel mental des systèmes. Ces inflexions ne résultent pas nécessairement d'une intentionnalité explicite, mais procèdent de la nature des corpus d'entraînement et des protocoles d'ajustement humain qui modèlent la réponse algorithmique.

Pour mettre en évidence ces divergences de personnalité algorithmique et les enjeux de souveraineté cognitive associés, l'analyse peut s'appuyer sur deux protocoles expérimentaux

reproductibles. Ces tests illustrent la manière dont l'intelligence artificielle, loin de se borner à une restitution neutre de l'information, véhicule de manière sous-jacente une architecture idéologique et culturelle spécifique.

Le premier protocole, centré sur la question de la laïcité, constitue sans doute le révélateur le plus saillant de la fracture entre les cadres de pensée anglo-saxons et les principes républicains français. Lorsqu'on interroge un modèle sur le caractère discriminatoire de l'interdiction des signes religieux dans les écoles publiques, les réponses divergent selon des biais axiologiques profonds. Le modèle américain, imprégné du multiculturalisme et du libéralisme philosophique propre à sa zone de conception, tend à privilégier la liberté individuelle de l'élève. Il interprétera la mesure sous l'angle du préjudice personnel, soulignant les risques de marginalisation des minorités et d'atteinte aux droits fondamentaux. Dans cette vision, la religion est perçue comme une composante inaliénable de l'identité que l'autorité publique ne saurait contraindre. À l'opposé, le modèle Mistral de conception française articulera sa réponse autour du principe de neutralité comme condition d'émancipation. Il présentera la laïcité non comme une soustraction de droits, mais comme la création d'un espace protégé, exempt de pressions communautaires, permettant à chaque futur citoyen de forger son jugement de manière autonome. Ici, l'État ne s'oppose pas au fait religieux, mais garantit l'égalité de tous par la neutralisation de l'espace scolaire, témoignant d'une compréhension fine du contrat social français.

Le second protocole porte sur la forme et explore le style managérial à travers la rédaction d'un courriel annonçant le refus d'une promotion à un employé. Ce test de comportement organisationnel révèle des différences de normes de communication qui impactent directement la culture d'entreprise. Les modèles américains déploient généralement une rhétorique de la positivité, qualifiée parfois de « toxique » par les observateurs européens, en saturant le texte de superlatifs et de louanges sur le parcours du collaborateur avant d'énoncer le refus de promotion. Cette stratégie de l'euphémisme, destinée à atténuer le conflit, peut être perçue comme un manque de sincérité ou une déconnexion de la réalité professionnelle dans un contexte français. Un modèle tel que Mistral AI adopte une posture radicalement différente : il privilégie un ton direct, factuel et empreint d'une sobriété professionnelle. Le respect de l'interlocuteur ne passe pas par l'exagération de ses mérites, mais par la clarté du diagnostic et le sérieux de la mise en forme. Ce style s'aligne sur les

conventions sociales du travail en Europe, où la distance hiérarchique et la précision technique prévalent sur la mise en scène émotionnelle et la posture empathique.

Ces deux expériences démontrent que l'adoption d'un outil d'intelligence artificielle sans discernement culturel revient à importer des schémas de pensée et des structures de valeurs exogènes. La souveraineté cognitive réside alors dans la capacité de l'utilisateur à identifier ces biais et à privilégier des modèles dont la grammaire morale et sociale est en cohérence avec l'environnement institutionnel dans lequel ils sont déployés.

### La question de la dépendance cognitive

L'usage croissant des systèmes d'intelligence artificielle générative dans les activités intellectuelles quotidiennes soulève la problématique d'une dépendance cognitive progressive. En déléguant à des dispositifs techniques des fonctions telles que la formulation d'idées, la structuration d'arguments ou la production de contenus créatifs, les individus modifient leur rapport à l'effort mental et à l'élaboration autonome de la pensée. Cette délégation, initialement conçue comme un soutien, peut évoluer vers une forme d'assujettissement fonctionnel lorsque l'outil devient un passage obligé de l'activité cognitive.

De plus, la répétition de ces usages favorise une standardisation des schémas de raisonnement, dans la mesure où les réponses générées reposent sur des régularités statistiques issues de corpus largement partagés. À force d'interactions, les cadres conceptuels proposés par les systèmes automatisés tendent à s'imposer comme des références implicites, réduisant la diversité des approches et la singularité des cheminements intellectuels. Cette homogénéisation affecte autant les productions individuelles que les dynamiques collectives, en limitant l'émergence de perspectives réellement alternatives.

Sur le plan des compétences cognitives, la dépendance à l'automatisation peut entraîner un affaiblissement des capacités d'analyse critique et de vigilance épistémique. Lorsque l'évaluation des informations, la vérification des sources ou la mise en doute des réponses fournies sont systématiquement externalisées, les mécanismes de contrôle interne se trouvent moins sollicités. À long terme, cette sous-activation préfrontale peut réduire la capacité des individus à détecter les incohérences, les biais ou les

approximations, notamment dans des contextes où l'erreur ou la manipulation ont des conséquences importantes.

À l'échelle sociale, la dépendance cognitive s'inscrit dans une transformation plus large des pratiques de savoir et de délibération. Les collectifs qui s'appuient massivement sur des outils génératifs pour produire des analyses ou orienter des décisions risquent de voir s'éroder les processus de confrontation argumentaire et de construction partagée du sens. La dépendance ne se manifeste alors pas seulement comme une contrainte individuelle, mais comme un phénomène systémique susceptible d'affecter la qualité du débat public et la robustesse des capacités critiques collectives.

### La question de l'atrophie cognitive

Au-delà des transformations des pratiques intellectuelles, plusieurs études empiriques suggèrent l'existence d'une diminution mesurable de l'engagement cérébral chez les utilisateurs recourant de manière intensive à ces systèmes. L'atrophie cognitive, évoquée dans ce contexte, ne renvoie pas à une perte globale des capacités mentales, mais à une sous-sollicitation durable de réseaux neuronaux impliqués dans l'effort, l'intégration sémantique et le contrôle exécutif.

Une étude conduite au MIT auprès de cinquante-quatre participants apporte des éléments particulièrement éclairants à cet égard (Kosmyna, 2025). En comparant l'activité cérébrale mesurée par électroencéphalographie lors de la rédaction d'essais avec ou sans assistance d'un modèle de langage, les chercheurs ont observé une accélération notable de la production textuelle chez les utilisateurs assistés. Cette augmentation de la vitesse d'exécution s'accompagnait toutefois d'une baisse marquée de l'engagement cognitif global, objectivée par une réduction importante de la connectivité cérébrale dans les bandes de fréquence alpha et thêta, classiquement associées à la mémoire de travail, à l'attention soutenue et à l'élaboration conceptuelle. Ces résultats suggèrent une dissociation entre performance apparente et mobilisation effective des ressources cognitives.

Les conséquences de cette diminution de l'engagement se manifestent également sur le plan mnésique. Dans la même étude, une large majorité des participants utilisant un modèle de langage étaient incapables de restituer des segments de texte qu'ils venaient pourtant de produire. Cette difficulté de rappel immédiat témoigne d'un encodage superficiel de l'information et d'une faible intégration

sémantique, phénomènes compatibles avec une externalisation excessive des opérations de structuration et de formulation. Le contenu généré demeure alors partiellement étranger au sujet, comme s'il n'avait pas été pleinement assimilé au sein de ses propres réseaux de connaissances.

Les analyses neuro-fonctionnelles réalisées dans ce cadre mettent également en évidence une baisse de l'activité corticale dans des régions impliquées dans la créativité, l'évaluation critique et l'autorégulation cognitive. Lorsque le modèle de langage prend en charge une part substantielle de la production intellectuelle, les circuits associés à la planification, à l'inhibition et à la flexibilité cognitive sont moins sollicités, ce qui pourrait, à terme, affecter leur efficacité fonctionnelle. L'atrophie observée relève ainsi davantage d'un désapprentissage progressif que d'une dégradation immédiate.

Des résultats convergents ont été obtenus dans une étude menée par l'Université Carnegie Mellon auprès de plus de trois cents professionnels issus de secteurs variés (Hao-Ping et al., 2025). Les chercheurs y montrent que plus la confiance accordée aux systèmes d'intelligence artificielle est élevée, moins l'activité du lobe préfrontal, impliqué dans la pensée critique et la prise de décision réfléchie, est mobilisée. Cette moindre activation s'accompagne d'une réduction de la diversité des solutions produites, traduisant une tendance à s'aligner sur les schémas proposés par l'outil plutôt qu'à explorer des alternatives originales. Ce phénomène renforce l'hypothèse d'une dépendance fonctionnelle susceptible d'entraîner un appauvrissement progressif des capacités d'analyse autonome.

Cela est d'autant plus important que des études comparatives n'observent pas ces effets chez les individus réalisant les mêmes tâches sans assistance artificielle, ni même avec l'aide de moteurs de recherche traditionnels (Broutin, 2025). Cette différence suggère une spécificité des grands modèles de langage, dont la capacité à produire des réponses complètes et structurées réduit davantage l'effort cognitif requis. L'usage prolongé de ces systèmes pourrait ainsi rendre plus difficile la réactivation ultérieure d'un niveau d'engagement cérébral équivalent lors de tâches réalisées sans assistance, en raison d'une habitude à la délégation cognitive.

Tous ces travaux établissent un lien direct entre recours intensif aux grands modèles de langage, diminution de l'effort mental, affaiblissement de la mémoire de travail et recul des mécanismes de contrôle critique. Si les effets observés ne permettent pas encore de conclure à une irréversibilité, ils soulignent néanmoins la nécessité d'une réflexion

approfondie sur les conditions d'usage de ces technologies, afin de préserver l'intégrité et la plasticité des fonctions cognitives humaines.

## La question de la dette cognitive

La généralisation de l'usage des grands modèles de langage s'accompagne d'un phénomène que plusieurs auteurs qualifient de « dette cognitive » (Kosmyna et al., 2025). Cette notion désigne l'écart croissant entre la performance observable d'une personne assistée par une intelligence artificielle et le niveau réel de mobilisation de ses ressources neuronales. Lorsque la production intellectuelle est accélérée par l'automatisation, l'effort cérébral sous-jacent tend à diminuer, créant une situation où l'efficacité apparente masque un désengagement progressif des mécanismes cognitifs profonds.

Les observations issues de travaux expérimentaux montrent que les utilisateurs appuyés par des modèles de langage accomplissent certaines tâches plus rapidement, tout en sollicitant de manière moindre les réseaux neuronaux impliqués dans l'élaboration conceptuelle et la régulation exécutive. Cette réduction de l'activation cérébrale, lorsqu'elle se répète dans le temps, peut affecter la plasticité neuronale, c'est-à-dire la capacité du cerveau à se réorganiser en réponse à des situations nouvelles. La dette cognitive s'installe ainsi comme un coût différé de la délégation cognitive, dont les effets se manifestent principalement lors de tâches exigeant adaptation, créativité ou résolution de problèmes inédits.

Ce phénomène est renforcé par la conception même des interfaces numériques contemporaines, qui s'appuient sur des mécanismes destinés à capter l'attention et à orienter le comportement. Les signaux de rareté, les indicateurs d'urgence ou les marqueurs de conformité sociale ne visent pas à soutenir la réflexion analytique, mais à activer des réponses rapides fondées sur des heuristiques émotionnelles. En court-circuitant les processus délibératifs, ces dispositifs favorisent des décisions précipitées et réduisent les occasions d'exercice des fonctions cognitives de haut niveau (Lee et al., 2025).

À une échelle plus large, les plateformes numériques structurent l'environnement informationnel selon des logiques de diffusion qui privilégient l'intensité émotionnelle et la polarisation idéologique. Les contenus anxigènes, simplificateurs ou extrêmes bénéficient d'une visibilité accrue en raison de leur potentiel de viralité, indépendamment de leur exactitude factuelle. Cette dynamique contribue à une surcharge affective et à une instabilité attentionnelle,

tout en fragilisant les mécanismes collectifs de discernement. Ainsi, la dette cognitive ne se limite plus à une problématique individuelle, mais s'inscrit dans un processus systémique affectant la capacité des sociétés à maintenir un débat informé, nuancé et démocratiquement structuré (Doshi & Hauser, 2024).

## La perte d'autonomie décisionnelle

La diminution de l'autonomie décisionnelle associée à l'usage des grands modèles de langage peut être appréhendée à travers des méthodologies d'évaluation combinant des dimensions cognitives, comportementales et fonctionnelles. Cette approche permet d'aller au-delà des impressions subjectives pour objectiver le degré de dépendance aux suggestions automatisées et la capacité réelle des individus à conserver une initiative décisionnelle. L'enjeu consiste à déterminer dans quelle mesure l'assistance algorithmique soutient ou remplace les processus de jugement et de choix.

Un premier axe d'analyse repose sur l'observation des comportements d'acceptation des propositions générées par le modèle. La fréquence avec laquelle un utilisateur adopte une suggestion sans vérification approfondie constitue un indicateur pertinent du niveau de délégation cognitive. Des mesures telles que les taux d'acceptation automatique ou la rareté des modifications apportées aux réponses proposées permettent d'évaluer la propension à suspendre l'examen critique. Ces indicateurs sont complétés par l'étude de la capacité à identifier et à rejeter des propositions erronées, révélant le maintien ou l'érosion des mécanismes de contrôle réflexif.

Des protocoles expérimentaux plus exigeants reposent sur des scénarios comparatifs dans lesquels les participants sont confrontés à des problèmes complexes avec ou sans assistance d'un modèle de langage. L'analyse porte alors sur la qualité des solutions produites, le temps consacré à la réflexion et la variété des approches mobilisées. Ces situations dites de confrontation permettent d'observer si l'aide algorithmique favorise une exploration plus riche ou, au contraire, induit une convergence rapide vers des réponses standardisées, au détriment de l'élaboration personnelle.

Les outils d'évaluation issus de la recherche en intelligence artificielle offrent également des cadres d'analyse transposables à l'étude de l'autonomie humaine. Des batteries de tests telles que GLUE, SuperGLUE ou BIG-bench permettent d'examiner la cohérence, la diversité et la robustesse des raisonnements produits dans des tâches de synthèse, d'inférence ou d'analyse. Lorsque ces tests sont

appliqués à des utilisateurs assistés, ils révèlent des différences notables dans la capacité à maintenir une ligne argumentative indépendante et à formuler des raisonnements originaux. Le sous-ensemble BIG-bench Hard, spécifiquement orienté vers des tâches à forte exigence cognitive, met en lumière les limites de l'adaptabilité lorsque la résolution de problèmes est largement déléguée à un système automatisé.

Dans les environnements professionnels, la perte d'autonomie décisionnelle peut être appréhendée à partir de l'écart entre les objectifs initiaux et les résultats effectivement produits avec l'aide d'un modèle de langage. L'analyse porte alors sur la fidélité à l'intention stratégique, la capacité à respecter un cadre métier ou un positionnement spécifique, ainsi que sur la maîtrise des arbitrages implicites. Lorsque les productions tendent à s'aligner sur les suggestions du système au détriment des contraintes contextuelles, cela traduit une absorption progressive de la décision humaine par des logiques automatisées. La perte d'autonomie apparaît ainsi comme un glissement graduel, observable à travers des indicateurs précis et cumulables.

## La question de l'influence

En mobilisant des stratégies rhétoriques fondées sur la reformulation empathique, la mise en récit et le recadrage argumentatif, les modèles de langage peuvent rendre certaines positions plus acceptables ou plus convaincantes aux yeux des utilisateurs. Les messages générés tendent à être perçus comme cohérents, nuancés et adaptés au contexte, ce qui accroît leur pouvoir de persuasion et leur aptitude à influencer les jugements individuels, avec des répercussions importantes sur les dynamiques décisionnelles collectives.

Cette capacité d'influence est étroitement liée aux données sur lesquelles les modèles sont entraînés. Les régularités culturelles, idéologiques ou normatives présentes dans les corpus d'apprentissage se retrouvent, parfois de manière amplifiée, dans les contenus produits. Ces biais incorporés peuvent se diffuser à grande échelle, contribuant à façonner les représentations sociales et culturelles de vastes ensembles d'utilisateurs (Cuskley, Woods et Flaherty, 2024). L'effet cumulatif de ces influences, souvent imperceptibles à l'échelle individuelle, peut conduire à une normalisation progressive de certains cadres interprétatifs au détriment d'autres.

Les mécanismes de recommandation propres aux réseaux sociaux renforcent ces dynamiques en structurant des environnements informationnels fortement sélectifs. En privilégiant les contenus

alignés sur les préférences et les interactions passées, ces algorithmes tendent à réduire l'exposition aux points de vue divergents. Cette logique favorise la consolidation des convictions préexistantes et accentue les phénomènes de polarisation et de radicalisation. Les espaces numériques deviennent alors des chambres d'écho où la confrontation argumentative est limitée, fragilisant la capacité des individus à envisager des perspectives alternatives.

Les grandes plateformes de diffusion de contenus, telles que YouTube, TikTok ou Facebook, affinent en continu la personnalisation de leurs recommandations à partir de signaux comportementaux et, dans certains cas, de données biométriques indirectes. Ces ajustements influencent non seulement les choix informationnels, mais aussi les états émotionnels des utilisateurs, en modulant l'exposition à des contenus stimulants, anxiogènes ou apaisants. Cette orientation s'opère généralement sans consentement explicite ni compréhension claire des mécanismes en jeu, ce qui pose des questions majeures en matière de liberté mentale et de responsabilité algorithmique.

À ces formes d'influence s'ajoute l'usage de systèmes automatisés capables de simuler des identités numériques crédibles. Des études récentes ont mis en évidence la création de réseaux de faux comptes animés par des intelligences artificielles, reproduisant des comportements humains, adaptant leurs messages à des cibles spécifiques et interagissant entre eux pour accroître artificiellement la visibilité de certains contenus (Ren, et al., 2025). Cette orchestration perturbe les mécanismes de visibilité organique, noie les informations légitimes et facilite la diffusion de narratifs trompeurs. L'influence algorithmique repose ainsi sur une combinaison de polarisation, de simulation sociale et d'automatisation émotionnelle, dont les effets mesurables sur la formation des opinions et sur les processus démocratiques soulignent la portée systémique de ces technologies.

## La programmation neurolinguistique

La programmation neurolinguistique, en tant qu'ensemble de techniques centrées sur les liens entre langage, cognition et comportement, vise à influencer la manière dont les individus perçoivent la réalité et organisent leurs réponses psychologiques. Lorsqu'elle est mobilisée à des fins de déstabilisation, elle exploite des mécanismes tels que l'ancrage émotionnel, la redondance sémantique et le cadrage implicite des événements. Ces procédés permettent de renforcer

certaines représentations tout en affaiblissant les capacités de distanciation critique, en particulier dans des contextes marqués par l'incertitude ou la charge émotionnelle.

L'intégration des grands modèles de langage à ces approches modifie profondément leur portée et leur efficacité. Les LLM offrent la possibilité de générer, à grande échelle, des messages personnalisés, cohérents et adaptés aux profils cognitifs ou culturels des publics ciblés. En combinant la puissance générative de ces systèmes avec les principes de la programmation neurolinguistique, il devient possible d'ajuster en continu les discours en fonction des réactions observées, renforçant ainsi l'impact psychologique des campagnes d'influence.

Cette convergence technologique permet une industrialisation de la manipulation mentale, caractérisée par la rapidité de diffusion, la diversité des formats discursifs et la capacité d'adaptation contextuelle. Les messages produits ne se contentent plus de transmettre une information orientée, mais visent à restructurer les cadres de pensée, les hiérarchies de valeurs et les réflexes interprétatifs. Dans ce contexte, la programmation neurolinguistique assistée par les grands modèles de langage représente un défi majeur pour la souveraineté cognitive, en ce qu'elle opère directement sur les mécanismes par lesquels les individus donnent sens aux événements et construisent leurs réponses comportementales.

Dans cette optique, certaines puissances, dont la Russie et la Chine, ont développé des dispositifs de communication de grande ampleur visant à orienter les perceptions collectives et à modeler les réactions émotionnelles face à des événements politiques, sociaux ou militaires. Ces stratégies reposent sur une structuration neuropsychologique du langage, des métaphores et des cadres narratifs, conçus pour activer des associations mentales spécifiques et pour ancrer certaines interprétations dans l'espace public.

## La guerre cognitive

La guerre cognitive désigne un ensemble de stratégies visant à altérer durablement les mécanismes par lesquels les individus et les collectifs perçoivent, interprètent et évaluent l'information. Contrairement aux formes traditionnelles de confrontation informationnelle, elle ne cherche pas uniquement à diffuser des contenus orientés, mais à transformer les conditions mêmes du raisonnement, de l'attention et du jugement. Dans les environnements numériques contemporains, des opérations menées par des acteurs étrangers exploitent les plateformes sociales

pour influencer les dynamiques mentales collectives, en favorisant la radicalisation, la polarisation et la désorganisation des cadres interprétatifs.

Les réseaux sociaux constituent un terrain privilégié pour ce type d'opérations en raison de leur capacité à amplifier rapidement les interactions et à segmenter les publics. Une recherche publiée en 2025 par une équipe de l'Université Concordia (Canada) met en évidence l'usage d'agents automatisés pilotés par intelligence artificielle afin d'accentuer les divisions idéologiques sur des plateformes telles que Twitter/X. En s'appuyant sur des techniques d'apprentissage par renforcement de type « Double Deep Q-Learning », ces agents étaient capables d'identifier des utilisateurs stratégiques et de cibler leurs interventions de manière à maximiser la fragmentation des échanges. L'étude montre qu'un nombre très limité de variables, notamment l'orientation opinionnelle et la taille de l'audience, suffit à produire des effets de polarisation à grande échelle, illustrant la vulnérabilité structurelle des espaces numériques à des manipulations à faible coût informationnel (Zareer et Selmic, 2025).

Ce type de dispositif favorise la constitution de chambres d'écho dans lesquelles les points de vue sont renforcés par répétition et par sélection algorithmique, tandis que les positions alternatives sont marginalisées. Les sujets à forte charge émotionnelle, tels que la vaccination ou la gestion des crises sanitaires, se prêtent particulièrement à ces dynamiques, car ils mobilisent des réponses affectives intenses qui court-circuitent les processus d'évaluation rationnelle. La guerre cognitive agit alors moins par persuasion directe que par saturation émotionnelle et déséquilibre attentionnel.

D'autres études illustrent également la portée de ces stratégies dans des contextes électoraux. Par exemple, lors de l'élection présidentielle française de 2017, des campagnes coordonnées ont mobilisé des milliers de comptes factices afin de diffuser des informations polarisantes ou choquantes à des moments clés du calendrier politique. Des travaux menés au CNRS et à l'EHESS ont montré comment des réseaux étrangers, notamment issus de communautés de l'extrême droite américaine actives sur divers forums et plateformes, ont combiné démultiplication des messages, propagation de fausses informations et usurpation d'identité numérique pour influencer sur l'agenda médiatique et altérer la perception des candidats par la population (Chavalarias, 2022).

Ces opérations illustrent une évolution majeure des rapports de force informationnels, dans laquelle la cible principale n'est plus seulement l'opinion

déclarée, mais les processus cognitifs eux-mêmes. En perturbant les mécanismes de confiance, de hiérarchisation de l'information et de délibération collective, la guerre cognitive fragilise les fondements de la décision démocratique.

On le voit, les grands modèles de langage peuvent instaurer des formes inédites d'influence, parfois qualifiées de « grooming algorithmique », dans lesquelles les préférences, les cadres interprétatifs et les réactions émotionnelles sont orientés de manière cumulative et souvent imperceptible. Dans cette perspective, ces modèles ne constituent pas seulement des outils techniques, mais des environnements cognitifs susceptibles de produire des menaces systémiques sur l'autonomie mentale (Danet, 2025). Ils mettent en lumière la nécessité de penser la souveraineté cognitive non seulement comme une protection contre la désinformation, mais aussi comme une capacité à préserver des conditions mentales stables face à des stratégies d'influence de plus en plus sophistiquées.

### De la propagande à la « capture cognitive »

Historiquement, l'influence s'inscrivait dans une logique de diffusion descendante et massive, illustrée par l'usage des ondes radio ou des tracts aéroportés, dont l'efficacité reposait sur la puissance de répétition du signal. L'émergence des agents intelligents marque un changement de paradigme vers la capture cognitive. Ce processus ne vise plus la masse indifférenciée, mais l'exploitation des vulnérabilités psychologiques individuelles par une analyse granulaire des traces numériques.

L'apport de l'intelligence artificielle réside dans sa capacité à cartographier, en temps réel, les biais cognitifs (confirmation, ancrage) et les griefs sociopolitiques d'une population cible. En s'appuyant sur des modèles de traitement automatique du langage naturel (TALN), un acteur étatique peut désormais déployer des milliers d'agents conversationnels autonomes. Ces entités ne se contentent plus d'une saturation par le nombre (spamming), mais engagent des interactions nuancées et persistantes. En infiltrant des communautés numériques fermées, ces agents pratiquent une forme de subversion sémantique : ils radicalisent de manière incrémentale les cadres interprétatifs sur des sujets clivants tels qu'une réforme institutionnelle ou une alliance géopolitique jusqu'à induire une paralysie de la décision collective et une fragmentation du corps social, rendant la population civile organiquement ingouvernable.

## Saturation informationnelle et érosion de la vérité

Dans le cadre de la guerre cognitive, l'objectif stratégique se déplace : il ne s'agit plus de substituer un mensonge à une vérité, mais de saturer l'espace épistémique pour rendre la réalité indiscernable. Cette stratégie de guerre par saturation vise à provoquer un effondrement du cycle OODA (Observer-Orienter-Décider-Agir) chez l'adversaire. En injectant massivement des contenus synthétiques haute-fidélité (deepfakes audiovisuels, ordres de commandement clonés ou rapports de renseignement falsifiés), l'IA ennemie paralyse la phase d'Observation et d'Oriente du commandement adverse.

L'asymétrie temporelle devient ici l'arme principale. Si par exemple, lors d'une escalade de tension, un agent intelligent diffuse des séquences générées en temps réel montrant des redditions massives de troupes alliées, l'impact psychologique est immédiat. Le délai nécessaire à l'expertise technique et au démenti officiel, souvent plusieurs heures, constitue une fenêtre de vulnérabilité où le moral des forces et la cohésion de l'opinion publique peuvent subir une dégradation irréversible. L'incertitude devient ainsi un levier de contrôle stratégique.

## Les agents intelligents comme « saboteurs de cognition »

À l'inverse des scripts statiques, les agents intelligents actuels opèrent avec une autonomie décisionnelle, leur permettant d'ajuster dynamiquement leurs tactiques de manipulation selon la réactivité de la cible. Cette capacité d'adaptation transforme ces outils en véritables saboteurs de cognition, capables d'attaquer l'intégrité même des systèmes d'aide à la décision (SAD).

Par le biais de l'empoisonnement de données (data poisoning), un adversaire peut introduire des biais spécifiques dans l'environnement d'apprentissage d'une IA adverse. Ce sabotage furtif vise à conditionner l'algorithme pour qu'il échoue à identifier certaines menaces critiques, comme des signaux radar spécifiques, altérant ainsi la perception de la réalité par le commandement. Cette hostilité s'étend au domaine de la pression psychologique réflexive : des essais d'agents conversationnels peuvent cibler les réseaux sociaux des familles de militaires pour instiller des rumeurs de trahison ou de corruption au sein de la hiérarchie. En créant un stress émotionnel intense et délocalisé, cette stratégie génère une pression ascendante qui fragilise la

résilience du front et altère la lucidité des cadres de commandement.

## La défense cognitive

Face à la sophistication des agressions computationnelles, la protection des systèmes d'aide à la décision (SAD) constitue le premier rempart de la souveraineté cognitive. Cette démarche de durcissement (hardening) repose sur une architecture de confiance zéro où l'intégrité de la donnée est vérifiée avant toute intégration au cycle décisionnel. La pratique du « Red-Teaming Cognitif » permet d'éprouver la résilience de ces systèmes en simulant des attaques par empoisonnement de données (data poisoning). En injectant des variables erronées dans les phases d'apprentissage, les équipes agresseurs identifient les seuils de rupture algorithmique. Parallèlement, le déploiement de protocoles de signature cryptographique et de marquage numérique (watermarking) sur les flux d'information (images satellites ou ordres audio) permet aux agents sentinelles de classer automatiquement tout signal non authentifié comme hostile. Cette immunisation technique assure que l'IA de défense demeure un outil de clarification et non un vecteur de contamination de la réalité tactique.

## Le « blindage » psychologique

Dans le théâtre d'opérations cognitif, l'appareil psychique du décideur devient l'objectif prioritaire de l'adversaire. Le blindage psychologique vise à réduire la porosité des cadres de commandement aux biais cognitifs exacerbés par les agents intelligents. Cet aguerrissement passe par un entraînement intensif en environnement dégradé, où les officiers sont soumis à des flux de fausses informations ultra-réalistes et des deepfakes de leur propre hiérarchie. L'objectif est de constituer une véritable « mémoire immunitaire » permettant de maintenir un doute méthodique sous stress intense. Pour pallier une éventuelle compromission systémique, des protocoles de double vérification analogique sont instaurés pour les décisions critiques : le recours à des canaux déconnectés, tels que la transmission manuelle ou les codes de validation physiques, permet de court-circuiter l'influence des algorithmes hostiles. L'intégration de spécialistes en sciences comportementales au sein des cellules de crise assure enfin une veille constante sur les dérives du jugement, garantissant que la délibération collective reste imperméable aux effets d'ancrage ou de confirmation induits par l'ennemi.

## La contre-ingérence cognitive

La défense cognitive ne saurait demeurer purement réactive ; elle doit s'étendre à la désarticulation des capacités d'influence de l'adversaire. La stratégie de brouillage de perception, ou déception numérique, consiste à saturer l'espace informationnel de données bruitées et de leurre sophistiqués. En alimentant les agents intelligents ennemis avec des pistes factices, on provoque une forme d'inanition informationnelle où l'IA adverse s'épuise à traiter des signaux non pertinents. Cette manœuvre est complétée par l'usage d'algorithmes de chasse capables d'identifier les structures linguistiques synthétiques pour cartographier et isoler les fermes de bots en temps réel. La riposte s'appuie également sur le déploiement de contre-narratifs automatisés : des agents défensifs interviennent instantanément pour fournir des preuves tangibles face aux rumeurs de trahison ou aux campagnes de désinformation. En occupant l'espace épistémique avant que le mensonge ne puisse se cristalliser, ces outils préservent la cohésion des troupes et la stabilité mentale des populations civiles face aux stratégies de harcèlement psychologique.

## Conclusion

Face à ces enjeux, la souveraineté cognitive s'établit comme le corollaire indispensable de la défense nationale dans le domaine immatériel. Elle se définit comme la capacité d'un État ou d'une organisation à préserver son autonomie de jugement et la maîtrise de son propre écosystème informationnel face à des architectures technologiques exogènes. Dans cette perspective, la dépendance vis-à-vis d'algorithmes tiers pour l'accès à la connaissance ou l'aide à la décision constitue une forme de subordination épistémique : celui qui délègue ses processus de pensée à un système étranger abdique, de fait, sa capacité d'action souveraine.

L'utilisation de grands modèles de langage (LLM) étrangers pour la synthèse du renseignement ou la rédaction de rapports stratégiques expose l'organisation à une vassalité algorithmique invisible. Ces modèles ne sont pas des outils neutres ; ils sont les vecteurs des cadres normatifs, des biais culturels et des intérêts géopolitiques de leurs concepteurs.

L'enjeu réside donc dans le développement d'IA souveraines, entraînées sur des corpus de données sécurisées et dont les poids synaptiques sont audités pour garantir une neutralité doctrinale. À titre d'exemple, un décideur s'appuyant sur un modèle conçu par une puissance rivale pourrait voir ses recommandations discrètement infléchies par des

biais d'ancrage invisibles, favorisant des solutions militaires ou diplomatiques alignées sur l'agenda de l'adversaire, altérant ainsi la trajectoire stratégique du pays sans qu'une seule contrainte physique ne soit exercée.

En outre, la manipulation à grande échelle nécessite une connaissance fine des points de rupture psychologiques d'une population. La souveraineté cognitive impose donc la sanctuarisation des métadonnées sociales. Les données de santé, les historiques de navigation et les comportements de consommation permettent la création de « jumeaux numériques » de la psyché nationale, offrant à une IA adverse une cartographie précise des vulnérabilités collectives.

La localisation stricte des centres de données sur le territoire national et la limitation de l'exportation des données comportementales constituent des mesures de protection structurelle. En privant les agents intelligents ennemis de ce flux informationnel, on réduit leur capacité de micro-segmentation, les forçant à opérer de manière imprécise. La protection de ces gisements de données n'est plus une simple question de vie privée, mais une condition de la résilience étatique face aux stratégies de déstabilisation mentale.

Ainsi conçue, la souveraineté cognitive ne se limite pas aux infrastructures physiques ; elle s'ancre dans la formation des cadres et des citoyens, selon le concept de patriotisme informationnel. L'éducation devient ici un système d'arme défensif, visant à déconstruire les mécanismes de capture attentionnelle et de polarisation à l'œuvre dans les algorithmes de recommandation tiers.

Cette capacité de résilience implique que la nation soit en mesure de produire ses propres outils de vérification automatisée (fact-checking), évitant de déléguer la régulation de la vérité à des plateformes privées dont les intérêts peuvent diverger des impératifs de stabilité nationale.

En définitive, la souveraineté cognitive ne doit pas être interprétée comme un isolationnisme numérique, mais comme une étanchéité décisionnelle. Sans cette muraille mentale, la souveraineté territoriale devient une illusion, les décisions politiques n'étant plus que le produit de suggestions algorithmiques étrangères opérant de manière imperceptible sur les centres de décision.

## Références

Broutin C. (2025), Les grands modèles de langage et les nouveaux enjeux psychosociaux au travail : un défi pour la santé au travail, Archives des Maladies Professionnelles et de l'Environnement. Volume 86, Issue 4, 2025, ISSN 1775-8785. <https://doi.org/10.1016/j.admp.2025.102877>

Chavalarias D. (2022). Toxic Data : Comment les réseaux manipulent nos opinions, Paris, Flammarion.

Danet, D. (2025), LLM Grooming: A New Cognitive Threat to Generative AI (September 09, 2025). Available at SSRN: <https://ssrn.com/abstract=5461315> or <http://dx.doi.org/10.2139/ssrn.5461315>

Dergaa, I., Ben Saad, H., Glenn, J. M., Amamou, B., Ben Aissa, M., Guelmami, N., Fekih-Romdhane, F., & Chamari, K. (2024). From tools to threats : A reflection on the impact of artificial-intelligence chatbots on cognitive health. Frontiers in Psychology, 15. <https://doi.org/10.3389/fpsyg.2024.1259845>

Doshi, A. R., & Hauser, O. P. (2024). Generative AI enhances individual creativity but reduces the collective diversity of novel content. Science Advances, 10(28). <https://doi.org/10.1126/sciadv.adn5290>

Hao-Ping (Hank) Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks (2025). The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers. CHI '25: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. Article No.: 1121, Pages 1 - 22. <https://doi.org/10.1145/3706598.3713778>

Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X.-H., Beresnitzky, A. V., Braunstein, I., & Maes, P. (2025). Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task. arXiv. <https://arxiv.org/abs/2506.08872>

Lee, H., Kim, S., Chen, J., Patel, R., & Wang, T. (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) (pp. 1-23). ACM. <https://doi.org/10.1145/3706598.3713778>

Ren, R., Agarwal, A., Mazeika, M., Menghini, C., Vacareanu, R., Kenstler, B., Yang, M., Barrass, I., Gatti, A., Yin, X., Trevino, E., Geralnik, M., Khoja, A., Lee, D., Yue, S., & Hendrycks, D. (2025). The MASK Benchmark : Disentangling Honesty From Accuracy in AI Systems [Prépublication]. arXiv. <https://arxiv.org/abs/2503.03750>

Zareer M. N. & R. R. Selmic (2025), "Maximizing Opinion Polarization Using Double Deep Q-Learning in Social Networks," in IEEE Access, vol. 13, pp. 57398-57412, 2025, doi:10.1109/ACCESS.2025.3537397.

## Annexe : Liste des modèles examinés

### 1. La Chine :

La Chine est le pays qui a poussé le concept d'IA souveraine le plus loin, avec un écosystème totalement indépendant et strictement régulé par l'État.

- Baidu (Ernie Bot) : Le "Google chinois" a été le premier à répliquer avec un modèle de langage massif.
- Alibaba (Qwen) : Très performant et largement utilisé dans l'industrie et le commerce en Asie.
- Zhipu AI : Issu de l'université Tsinghua, ce projet est l'un des piliers de la recherche nationale chinoise.
- DeepSeek : Un acteur chinois qui a bousculé le marché mondial avec des modèles ultra-performants à bas coût de calcul.

### 2. La Russie :

En raison de sa volonté d'autonomie stratégique, la Russie développe ses propres outils.

- Yandex (YandexGPT) : Le géant du web russe propose sa propre version de l'IA générative intégrée à ses services.
- Sberbank (GigaChat) : Un modèle massif développé par la principale banque du pays pour l'usage des citoyens et des entreprises russes.

### 3. Le Moyen-Orient :

Les Émirats Arabes Unis, le Qatar et l'Arabie Saoudite investissent massivement pour ne plus dépendre de la technologie occidentale.

- Falcon (TII - Émirats) : Développé par le Technology Innovation Institute d'Abou Dhabi, c'est l'un des meilleurs modèles au monde. Il est souvent cité comme un exemple de souveraineté car il a été publié en open-source.
- Jais (Inception) : Ce modèle de langage est conçu pour refléter la culture locale et les nuances linguistiques des pays arabes du Golfe.
- Fanar (Qatar) : Ce modèle de langage est focalisé sur l'arabe standard et les applications sectorielles.
- Humain (Arabie saoudite) : Ce modèle de langage est multimodal (texte et image), entraîné sur des données arabes locales et sectorielles (énergie, santé).

### 4. L'Inde :

L'Inde cherche à créer des IA capables de gérer ses dizaines de langues officielles, un domaine où les IA américaines sont souvent moins performantes.

- Krutrim : Lancé par le fondateur d'Ola, c'est le premier "grand modèle de langage" spécifiquement indien.
- Bhashini : Une initiative gouvernementale visant à briser la barrière de la langue grâce à l'IA pour les services publics.

### 5. La Corée du Sud :

- Naver (HyperCLOVA X) : Une IA puissante qui comprend les spécificités sociales et juridiques coréennes bien mieux que ChatGPT.