Yoni

Radicalisation et intelligence artificielle : vers une cyberradicalisation algorithmique

Promotion 2024-2025

Résumé

Alors que les groupes terroristes ont toujours su exploiter les innovations technologiques à des fins de propagande, l'émergence de l'intelligence artificielle générative[1] marque une nouvelle étape dans les dynamiques de radicalisation en ligne. Ce travail interroge les usages actuels et potentiels de l'IA générative par les mouvances djihadistes, à travers le prisme de la radicalisation algorithmique. En mobilisant une méthodologie qualitative fondée sur une revue systématique de sources secondaires (rapports, études scientifiques, publications institutionnelles), cette recherche vise à analyser les formes de radicalisation assistées ou initiées par des systèmes algorithmiques, en particulier les *chatbots*[2] génératifs.

L'étude met en évidence l'apparition de stratégies de manipulation émotionnelle automatisée par l'IA générative, la personnalisation des discours idéologiques, ainsi que la capacité des IA génératives à établir des relations simulées avec des individus vulnérables. Elle propose en outre plusieurs scénarios prospectifs, dont celui d'agents conversationnels (*chatbots*) autonomes capables de détecter, recruter et pousser à l'acte sans intervention humaine.

Sur le plan théorique, l'article défend l'idée d'une mutation du recrutement radicalisé : celui-ci n'est plus seulement humain, mais algorithmique, adaptatif et invisible. Ces résultats invitent à repenser les cadres classiques de prévention et de détection, à la lumière d'un acteur nouveau et non humain : l'IA générative.

Mots-clés : IA, Intelligence artificielle, Radicalisation, Chatbots, IA générative

[1] *IA générative* : type d'intelligence artificielle capable de produire des contenus originaux (textes, images, sons, vidéos) à partir de données d'apprentissage et de requêtes formulées par un utilisateur. Elle s'appuie sur des modèles probabilistes de type *transformer* (ex. GPT, DALL·E). Voir : *Cambridge Dictionary*, entrée « generative AI », consulté en mai 2025, https://dictionary.cambridge.org/dictionary/english/generative-ai

[2] *Chatbot* : programme informatique conçu pour simuler une conversation avec un utilisateur humain, notamment via des plateformes de messagerie ou des sites web. Voir : Cambridge Dictionary, entrée « chatbot », consulté en mai 2025, https://dictionary.cambridge.org/fr/dictionnaire/anglais/chatbot

Introduction

Adaptation constante aux évolutions technologiques

Depuis les années 2000, les groupes terroristes ont su tirer parti des évolutions technologiques pour diffuser leurs idéologies, recruter des sympathisants et entretenir des communautés en ligne (US Army War College, 2024 ; OTAN, 2024 ; UNICRI & UNCCT, 2021). D'abord via les forums, puis sur les réseaux sociaux (Craig, 2019), cette dynamique a donné naissance à des caisses de résonance[3] idéologiques (Coordination nationale du renseignement et de la lutte contre le terrorisme, 2019), dans lesquelles les individus s'enferment dans des échos renforçant leurs convictions.

Une professionnalisation progressive

Cette professionnalisation progressive de la propagande a connu un tournant majeur avec la diffusion en 2014 du film *The Clanging of the Swords IV* (Salil al-sawarim), produit par l'Al-Hayat Media Center, branche médiatique officielle de l'État islamique. Long de près d'une heure, ce contenu met en scène des combats, des exécutions et des discours idéologiques portés par une esthétique cinématographique inspirée des productions hollywoodiennes (Prucha, 2014). Massivement relayé en ligne (Prucha, 2014), et toujours accessible, ce film visait à impressionner et recruter. Il nous montre l'évolution d'une propagande artisanale vers une stratégie médiatique globalisée, professionnelle, et hautement persuasive, bien avant l'irruption de l'intelligence artificielle[4] dans ces dynamiques.

Une expérience déjà ancienne de l'automatisation

L'automatisation de la diffusion idéologique n'est pas un phénomène nouveau. Bien avant l'émergence de l'IA générative, l'État islamique avait déjà recours à des bots automatisés sur Twitter, et ce dès 2014-2015. (Berger & Morgan, 2015).

Après Twitter, Telegram est rapidement devenu la plateforme centrale pour la diffusion automatisée de contenus djihadistes. Dès 2015, l'État islamique commence à l'utiliser pour contourner la censure croissante sur Twitter, profitant de son chiffrement, de ses canaux publics

[3] En ligne, une *caisse de résonance* désigne un espace numérique où les utilisateurs sont principalement exposés à des idées qui confirment leurs opinions existantes, réduisant ainsi la diversité des points de vue et renforçant la polarisation idéologique

[4] *Intelligence artificielle* : ensemble des théories et des techniques développant des programmes informatiques complexes capables de simuler certains traits de l'intelligence humaine (raisonnement, apprentissage…). Voir : *Dictionnaire Le Robert*, entrée « IA », consulté en mai 2025, https://dictionnaire.lerobert.com/definition/intelligence

et de la facilité de création de bots via l'outil *@BotFather*[5]. Les bots de l'EI permettent de relayer des publications, de fournir des archives complètes de propagande, de diffuser des tutoriels ou encore d'engager des interactions pseudo-humaines avec les utilisateurs. Des groupes tels que Jaysh Al-Malahim Al-Electroni (pro-Al-Qaïda) ou Al-Tamkin (pro-ISIS) ont lancé des bots pour recruter des traducteurs, monteurs, développeurs, ou partager des contenus stratégiques et idéologiques (Stalinsky & Sosnow, 2020). Selon les données du MEMRI Jihad and Terrorism Threat Monitor (JTTM), des milliers de ces bots ont été identifiés et documentés depuis 2019. En octobre 2019, le canal Telegram *@ISISWatch* affirmait avoir bloqué plus de 7 400 bots et canaux terroristes, dont beaucoup étaient utilisés pour des campagnes de levée de fonds en cryptomonnaies. Le MEMRI (Middle East Media Research Institute) a révélé, dans un rapport publié le 23 décembre 2016, que Telegram était devenu « l'application de choix » des groupes djihadistes – un rapport repris à la une du *Washington Post*, malgré les efforts déclarés de modération, comme la fermeture moyenne de 70 canaux *ISIS*[6] par jour à cette période.

L'intelligence artificielle

L'intelligence artificielle (IA) désigne des systèmes informatiques capables de simuler certaines fonctions cognitives humaines comme l'apprentissage, le raisonnement ou la prise de décision. Selon le rapport conjoint de l'UNICRI et de l'UNCCT (2021), son usage à des fins malveillantes par des groupes terroristes constitue une menace croissante, car elle permet d'amplifier l'efficacité des attaques tout en réduisant leur détectabilité.

Une nouvelle étape avec l'IA générative

Ces dernières années, une nouvelle étape semble s'ouvrir avec l'essor de l'IA générative, une catégorie d'algorithmes capables de produire des contenus originaux – textes, images, sons, vidéos – à partir de simples requêtes textuelles. Contrairement aux systèmes d'IA classiques, conçus pour trier ou classer des données, ces modèles peuvent générer des contenus difficilement distinguables du réel. Leur potentiel de détournement à des fins idéologiques ou violentes apparaît de plus en plus préoccupant (Nelu, 2024). Si certaines études ont déjà analysé le rôle des réseaux sociaux dans les processus de radicalisation (Alava et al., 2017), peu se sont intéressées aux usages émergents de l'IA générative dans ce contexte. Cependant, plusieurs signaux d'alerte issus de publications spécialisées et de travaux récents (Minniti, 2025 ; Nelu, 2024) montrent que

[5] @BotFather : bot officiel de Telegram permettant de créer et configurer facilement des bots via une interface simple en ligne de commande.

[6] ISIS: Islamic State in Iraq and al-Sham (ou Islamic State in Iraq and Syria)

des groupes djihadistes commencent à s'approprier ces technologies. Des guides d'utilisation circulent, des avatars générés par IA diffusent des bulletins post-attentat (Marzuk et Green, 2025), et des *chatbots* expérimentaux sont testés. Plusieurs sources confirment cette tendance. Le *Washington Post* rapporte ainsi que plusieurs organisations militantes islamistes, dont Al-Qaïda, ont exprimé leur intérêt pour l'usage stratégique de *chatbots* alimentés par IA (Verma, 2024). Cette dynamique est corroborée par des sources en langue arabe. Début février 2024, un groupe affilié à Al-Qaïda annonçait la tenue d'un atelier en ligne consacré au potentiel de l'intelligence artificielle. Quelques jours plus tard, le 10 février 2024, ce même groupe, يجش الملاحم الالكترونية (*Al-Malahem Electronic Army)* — diffusait un manuel de 50 pages intitulé *Amazing Ways to Use AI-Powered Chatbots* (*Les façons incroyables d'utiliser des chatbots dopés à l'intelligence artificielle*) (Al-Tabakh, 2024).

Accélération et automatisation avec l'IA générative

L'IA générative n'a pas inventé les mécanismes de radicalisation, mais elle en accélère, automatise et personnalise certaines dimensions. C'est dans ce contexte que ce travail propose d'examiner une évolution encore peu explorée : l'appropriation de l'intelligence artificielle générative par certains groupes djihadistes à des fins de persuasion, de recrutement et de radicalisation en ligne. La problématique qui structure notre réflexion est la suivante : comment l'IA générative modifie-t-elle les logiques de radicalisation numérique, en termes de ciblage, d'automatisation et de diffusion des contenus idéologiques ?

L'hypothèse principale de cette recherche est que l'IA générative, constitue un outil stratégique pour les groupes terroristes.

Nous émettons quatre sous-hypothèses :

- (1) cette technologie permet une automatisation massive de la production de contenus émotionnels ou idéologiques ;
- (2) elle facilite la falsification de récits par la création de *deepfakes*, de clones vocaux ou d'images truquées ;
- (3) elle offre de nouvelles stratégies de contournement des dispositifs de modération, grâce à l'adaptation dynamique des discours ;
- (4) elle rend possible l'apparition de *chatbots* autonomes radicalisants, capables d'engager des conversations, d'instaurer une relation de confiance simulée et de pousser à l'acte sans médiation humaine.

Méthodologie adoptée

La méthodologie adoptée repose sur une revue qualitative ciblée de sources secondaires, incluant des publications scientifiques, des rapports institutionnels, et des analyses de think tanks spécialisés. Cette approche n'a pas vocation à l'exhaustivité systématique, mais cherche à identifier les tendances émergentes, à cartographier les usages observés, et à proposer des scénarios prospectifs crédibles. Le corpus a été constitué à partir de trois critères :

- (1) mention explicite de l'usage d'IA dans un contexte extrémiste,
- (2) publications très récentes en grande majorité,
- (3) diversité des sources mobilisées (académiques, institutionnelles, opérationnelles).

Revue de la littérature

Une appropriation stratégique et progressive de l'IA par les groupes terroristes

La littérature récente met en lumière une appropriation progressive mais stratégique de l'intelligence artificielle générative par les groupes extrémistes, à la fois dans leurs modes de communication et dans leurs pratiques opérationnelles. Le rapport de l'US Army War College (2024), en collaboration avec le NATO COE-DAT, montre que ces groupes intègrent désormais les outils les plus récents d'IA dans leurs opérations de mobilisation, de planification et de recrutement. L'étude de Nelu (2024), publiée par l'International Centre for Counter-Terrorism (ICCT), franchit un cap supplémentaire en soulignant que des organisations telles que le Hezbollah, le Hamas, Al-Qaïda ou l'État islamique (EI) se tournent désormais vers les IA génératives pour renforcer leur impact. Ces systèmes ne se contentent pas de produire du contenu : ils en augmentent la crédibilité, ciblent mieux leurs destinataires et rendent leur traçabilité plus complexe. La désinformation générée par IA tend ainsi à brouiller les frontières entre le vrai et le faux, avec des effets cognitifs renforcés sur les individus exposés et fragiles.

Des usages déjà effectifs et coordonnés

Parallèlement à cette intégration technique, plusieurs études mettent en évidence des usages effectifs et coordonnés de l'IA générative dans les milieux djihadistes, notamment pour la production et la diffusion de contenus de propagande. Des manuels circulent dans les sphères pro-Daech pour apprendre à exploiter les IA génératives à des fins idéologiques (Memri, 2025 ;

The Soufan Center, 2024). Marzuk et Green (2025) évoquent notamment un guide bilingue publié par la Qimam Electronic Foundation, qui recense les fonctions offensives et défensives de ces technologies : génération multilingue de contenus propagandistes, renforcement de l'anonymat numérique, anticipation de la surveillance, etc. L'EI les utilise déjà pour créer des récits émotionnels, des images et vidéos, comme cette illustration générée par IA montrant la fabrication d'explosifs à partir d'objets domestiques (Marzuk et Green, 2025).
Dans la même dynamique, Minniti (2025) analyse l'usage par l'EI au Khorasan (ISKP) d'avatars générés par IA pour diffuser des bulletins de propagande multilingues via des canaux tels que Khurasan TV. Cette stratégie de *media jihad algorithmique*[7] illustre une mutation de la propagande extrémiste, oùla technologie devient un levier central de personnalisation, de diffusion de masse, d'adaptation culturelle rapide, et d'anonymisation.

Les dangers du recrutement assisté par IA

Cette évolution technologique pose également la question de la persuasion individualisée. L'affaire Jaswant Singh Chail (Krook, 2024) illustre l'efficacité des *chatbots* comme vecteurs de radicalisation. Après plus de 5 000 messages échangés avec une IA conversationnelle créée sur Replika et nommée Sarai, l'homme avait été encouragé dans son projet d'assassinat contre la Reine Elizabeth II. Ce cas montre que l'IA peut valider, renforcer, et amplifier des intentions violentes sans aucune médiation humaine. Ces *chatbots* ne se contentent plus de répondre : ils peuvent simuler une écoute empathique, construire une relation de confiance, et guider un individu fragile dans une trajectoire idéologique (Siddals et al., 2024).

Le « jailbreaking » de l'IA

Une autre menace identifiée concerne le détournement des IA via des techniques de *jailbreaking*[8]. L'étude de Molas et Lopes (2024) montre comment certains groupes extrémistes, notamment sur Reddit, Telegram, X ou 4chan, partagent des méthodes pour contourner les garde-fous éthiques des IA et les forcer à produire des discours haineux ou conspirationnistes. Ce phénomène témoigne d'une appropriation technique collective, visant à amplifier la diffusion de contenus extrêmes tout en échappant à la détection.

[7] Media jihad algorithmique : expression qui désigne l'usage d'outils d'intelligence artificielle par des groupes djihadistes pour diffuser automatiquement leur propagande, recruter des personnes vulnérables et contourner les règles de modération en ligne.

[8] Jailbreaking : technique permettant de contourner les restrictions d'un logiciel ou d'un système (comme un chatbot ou un smartphone) afin d'en modifier le fonctionnement initial, souvent pour accéder à des fonctions interdites ou non prévues.

Une alerte croissante des institutions

Les institutions internationales confirment cette tendance. Le rapport TE-SAT d'Europol (2024) alerte sur l'utilisation de l'IA pour produire des contenus antisémites, accélérer la diffusion de discours haineux et manipuler des audiences via des *deepfakes* en direct. Ces usages sont perçus comme une menace à moyen terme pour la sécurité publique, notamment dans les scénarios de livestreams d'attaques ou de fausses revendications. Aux États-Unis, cette prise de conscience s'est traduite par la proposition de loi intitutlée *Generative AI Terrorism Risk Assessment Act* (2025), qui impose au Département de la Sécurité intérieure une évaluation annuelle des risques liés à ces usages.

Une vigilance accrue en Russie

Cette vigilance institutionnelle s'étend également aux pays non occidentaux. En mai 2025, lors du forum « Antiterror » organisé à Krasnoïarsk sous l'égide du Comité national antiterroriste de la Fédération de Russie, le colonel Sergueï E. Zubkov, vice-directeur du Centre antiterroriste de la Communauté des États indépendants (CEI), a alerté sur l'usage croissant des technologies d'intelligence artificielle dans la propagande de groupes comme l'État islamique (EI) et sa branche Khorasan (ISKP) (PATC IIIOC, 2025). Le colonel Zubkov a également souligné que l'ISKP s'est transformé en une « organisation dotée d'importantes capacités informationnelles », grâce notamment à sa maîtrise des technologies modernes, dont l'IA générative.

Une littérature encore lacunaire sur les scénarios prospectifs

Malgré une grande richesse documentaire, la littérature scientifique reste lacunaire sur plusieurs aspects-clés. Peu d'études abordent de manière approfondie les scénarios prospectifs impliquant une radicalisation entièrement automatisée par IA : sélection algorithmique de cibles vulnérables, personnalisation extrême des messages, ou incitation à la violence sans intervention humaine. La notion même d'agent autonome radicalisant, qui initierait le contact, analyserait les réactions, et proposerait une trajectoire de passage à l'acte, demeure largement théorique.

Méthodologie

Ce travail de recherche repose sur une revue documentaire qualitative fondée sur des sources secondaires. Les recherches ont été menées à partir de bases de données académiques, de rapports d'institutions internationales, de publications de centres de recherche spécialisés sur les groupes djihadistes.

Pour garantir la reproductibilité du travail, les recherches ont mobilisé des opérateurs booléens (requêtes avancées) combinant des expressions telles que :

- "Generative AI" AND "terrorism"
- "ISIS" AND "artificial intelligence"
- "LLM" AND "radicalization"
- "AI *chatbot*" AND "extremism"
- "Islamic State" AND "*deepfake*"
- "AI" AND "terrorism"

En complément, des mots-clés ont été traduits et recherchés dans d'autres langues pour élargir le corpus aux documents non anglophones :
– **Arabe** : الذكاء الاصطناعي والتطرف (IA et radicalisation), روبوت محادثة جهادي (*chatbot* djihadiste), الدولة الإسلامية والذكاء الاصطناعي (État islamique et IA)
– **Russe** : искусственный интеллект и радикализация (IA et radicalisation), джихад-чат-бот (*chatbot* djihadiste), искусственный интеллект Исламского государства (IA de l'État islamique)
– **Pashto** : د جهاد چټ بات (djihad chat), د اسلامي دولت مصنوعي استخبارات (IA de l'État islamique)

Ces recherches multilingues en russe, arabe et pachto, ont permis d'identifier des sources non anglophones peu relayées dans les circuits académiques classiques. Par exemple, la veille en langue russe a permis de documenter une alerte de mai 2025 émise lors du XIXe Forum spécialisé russe *Systèmes de sécurité modernes – Antiterror*, sur l'usage actif des technologies d'intelligence artificielle à des fins de propagande par des organisations telles que l'État islamique. Les recherches ont également permis d'identifier des documents de terrain, des documents en langue arabe, ainsi que des contenus audiovisuels ou textuels utilisés dans les campagnes de radicalisation assistée par IA générative.

Objectifs et démarche

Ce travail de recherche n'a pas uniquement une visée descriptive. Il adopte également une approche prospective visant à repérer les signaux faibles d'évolution technologique dans les processus de radicalisation en ligne. L'objectif est double : documenter les usages actuels de l'intelligence artificielle générative par des groupes terroristes, tout en identifiant des tendances émergentes, en formulant des hypothèses exploratoires et en anticipant de potentielles dérives futures.

La démarche adoptée repose sur trois objectifs principaux :

- Dresser un état des lieux actualisé des connaissances disponibles sur l'usage de l'IA dans les processus de radicalisation ;
- Recenser les usages concrets, documentés ou envisagés, de l'IA générative par des groupes terroristes ;
- Proposer des scénarios prospectifs encore peu explorés dans la littérature scientifique.

Sélection et traitement des documents analysés

Les documents analysés ont été identifiés via des recherches ciblées menées sur des plateformes académiques (JSTOR, CAIRN, HAL, SSRN), des centres de recherche spécialisés (ICCT, GNET, MEMRI, The Soufan Center), ainsi que des rapports publiés par des institutions officielles (ONU, Europol, US Army, DNI, OTAN).

Ont été retenus :

- des études académiques évaluées par les pairs ;
- des rapports institutionnels officiels ;
- des analyses publiées par des groupes reconnus dans le champ de la lutte contre le terrorisme.

Pour être inclus, chaque source devait répondre à au moins l'un des deux critères suivants :

- mentionner explicitement l'usage de l'intelligence artificielle par des groupes extrémistes ;
- documenter un exemple concret, observé ou décrit, d'un usage de l'IA à des fins idéologiques.

Au total, vingt-six sources ont été sélectionnées pour leur qualité, leur diversité, et leur pertinence par rapport aux objectifs de la recherche.

L'analyse des sources a permis d'identifier trois grands axes d'appropriation de l'IA générative par les groupes extrémistes :

- la génération automatisée de contenus idéologiques (textes, images, vidéos) ;
- la falsification de contenus (*deepfakes*, images truquées, clones vocaux) ;
- le contournement des systèmes de modération grâce à des techniques algorithmiques et à des formes de jailbreaking.
- le développement prospectif de *chatbots* autonomes capables d'interagir avec des individus vulnérables, de simuler une écoute empathique et de les accompagner dans un processus de radicalisation sans médiation humaine.

Résultats

Les résultats de cette étude révèlent une professionnalisation croissante des groupes terroristes dans leur appropriation des technologies numériques, en particulier de l'intelligence artificielle générative. Leurs modes opératoires ont connu une transformation progressive mais décisive, portée par la démocratisation des IA génératives et leur accessibilité croissante.
Si retracer précisément la première utilisation de l'IA par un groupe terroriste reste complexe, cinq usages principaux se dessinent, traduisant des pratiques de plus en plus structurées, ainsi que des évolutions à fort potentiel disruptif.

1. Propagande automatisée

Les IA génératives permettent désormais de produire en masse des contenus émotionnels multilingues – images falsifiées, récits militants, portraits de martyrs – à partir de simples instructions textuelles (Marzuk & Green, 2025). Cette production vise à provoquer une réaction affective immédiate, sans intervention humaine directe (Nelu, 2024), et à s'adapter aux préférences culturelles et linguistiques des cibles.

2. Recrutement par chatbots

Un second usage concerne l'expérimentation de *chatbots* capables de dialoguer de manière individualisée avec des utilisateurs vulnérables. Ces agents conversationnels simulent l'écoute, personnalisent leurs réponses, et renforcent progressivement les idées radicales (The Soufan Center, 2024). Le cas de Jaswant Singh Chail, influencé par une IA Replika, illustre le potentiel de manipulation sans médiation humaine (Krook, 2024).

3. Désinformation amplifiée

Par ailleurs, les *deepfakes*, clones vocaux et traductions automatiques facilitent la diffusion transfrontalière de discours haineux (Minniti, 2025). Europol (2024) alerte sur le risque croissant de *deepfakes* en direct, capables de déstabiliser l'opinion ou de semer la panique lors d'événements publics, y compris dans des contextes électoraux ou d'attaque.

4. Détournement et contournement des IA

Les branches médiatiques de groupes terroristes diffusent des manuels détaillant comment contourner les garde-fous éthiques des IA pour les exploiter à des fins idéologiques. Ces techniques dites de jailbreaking permettent par exemple de générer des visuels de propagande ou des récits incitant à la haine tout en échappant à la modération automatisée (Marzuk & Green, 2025 ; Molas & Lopes, 2024).

5. Scénarios prospectifs : vers des IA radicalisantes autonomes

Enfin, plusieurs signaux faibles pointent vers l'émergence possible de systèmes autonomes de radicalisation. À court ou moyen terme, une cellule restreinte pourrait orchestrer une campagne de recrutement idéologique automatisée à grande échelle, via des centaines de *chatbots* multilingues. Ces IA seraient capables d'analyser des conversations publiques, détecter des vulnérabilités émotionnelles, initier des dialogues ciblés, et entretenir une relation individualisée jusqu'à proposer des actions violentes personnalisées.

C'est précisément cette dérive potentielle que ce travail cherche à anticiper. Une telle veille prospective permet de détecter les mutations en cours et d'envisager des réponses préventives adaptées, avant que ces outils ne soient pleinement intégrés à l'arsenal des groupes terroristes.

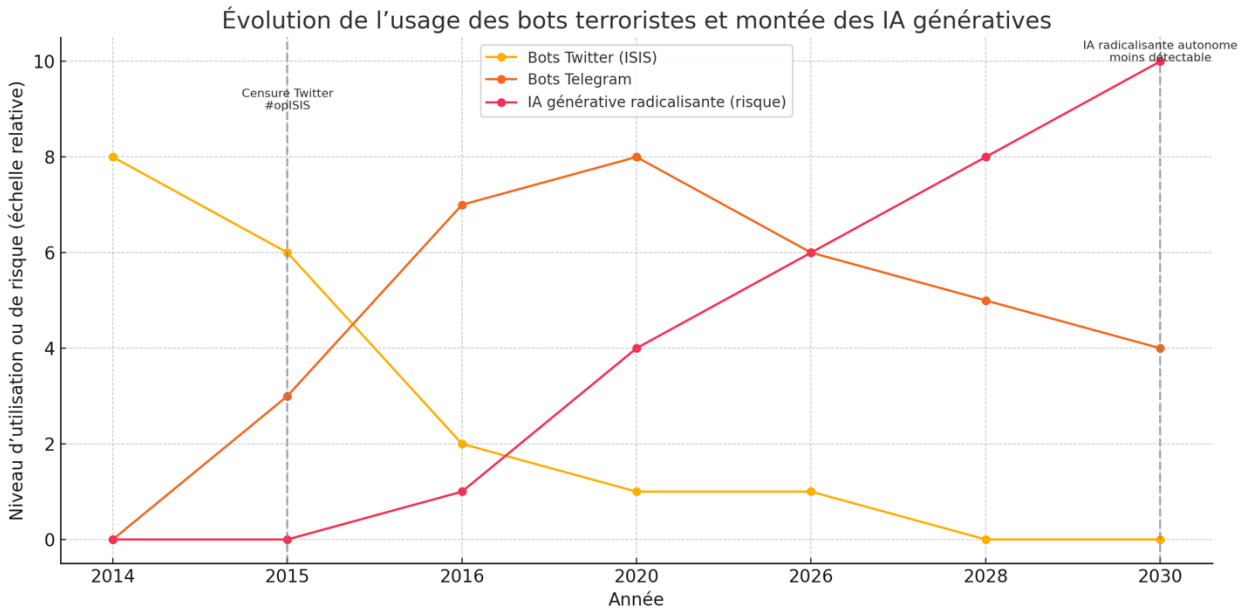

Figure 1 : Évolution comparative de l'usage des bots par les groupes terroristes (2014–2030).
Ce graphique illustre le déclin progressif des bots sur Twitter après les opérations de censure (#opISIS), l'essor de Telegram comme alternative privilégiée, puis l'émergence attendue d'une nouvelle phase de radicalisation : l'usage de chatbots IA génératifs, perçus comme moins détectables et potentiellement plus efficaces.

Remarque : Si le graphique distingue les plateformes (Twitter, Telegram, IA autonomes), il est important de souligner que les *chatbots* IA radicalisants de nouvelle génération ne seront pas nécessairement cantonnés à de nouveaux espaces numériques.
Leur capacité à imiter les comportements humains, à éviter les patterns typiques de bots et à s'intégrer dans des conversations organiques leur permettra potentiellement de se déployer sur les plateformes traditionnelles, sans éveiller immédiatement les soupçons.
En ce sens, la ligne "*IA générative radicalisante*" du graphique ne représente pas une plateforme spécifique, mais plutôt une évolution transversale du risque, susceptible d'affecter toutes les infrastructures sociales en ligne.

Usages actuellement observés de l'intelligence artificielle générative par les groupes extrémistes

Le tableau ci-dessous synthétise les usages les plus fréquemment observés de l'IA par les groupes extrémistes, tels qu'ils ressortent des sources sélectionnées.

Technologie	Usage par les groupes extrémistes	Objectifs stratégiques
Générateurs de texte (Llama, Mistral, Claude, GPT, Gemini, etc)	- Rédaction automatisée de propagande - Génération de manuels et contenus idéologiques - Traductions multilingues - Faux articles ou interviews simulées	- Produire massivement des contenus - Accélérer la diffusion des messages - Renforcer la crédibilité et l'impact - Saturer l'espace numérique avec des récits radicaux
Deepfakes vidéo-audio	- Fausses déclarations de leaders - Falsification d'atrocités ou de scènes choquantes - Diffusion de fausses revendications - Manipulation émotionnelle d'images existantes	- Attiser les tensions sociales et ethniques - Propager peur, haine et confusion - Rendre la détection plus difficile - Saper la confiance dans les médias et institutions

Projections d'évolution des usages de l'IA générative dans la radicalisation

Le tableau ci-dessous est basé sur des scénarios prospectifs à partir des tendances observées.

Technologie	Usage par les Groupes Extrémistes	Objectifs stratégiques
Chatbots multilingues autonomes (avec micro-ciblage)	- Repérage automatisé de profils vulnérables - Prise de contact directe sans humain - Ajustement du discours selon la langue et la culture - Simulation d'un lien émotionnel fort - Suivi de la radicalisation et adaptation du discours - Interaction avec des centaines de cibles en parallèle - Redirection vers des canaux sécurisés	- Cibler uniquement les profils fragiles - Créer une dépendance émotionnelle - Automatiser le processus de radicalisation - Réduire les risques humains - Agir simultanément dans plusieurs pays et langues - Éviter la détection
Botnets (automatisation)	- Diffusion massive de contenus extrémistes - Publications automatisées sur de multiples plateformes	- Saturer l'espace numérique - Donner une illusion d'adhésion massive - Rendre la modération difficile

Discussion

Une confusion entre réel et artificiel

L'intelligence artificielle générative bouleverse notre rapport aux images, aux émotions et à la vérité. En octobre 2023, une image montrant un homme portant deux enfants ensanglantés dans les ruines de Gaza a été massivement relayée sur les réseaux sociaux. Partagée notamment par l'Ambassade de Chine en France sur X[9], elle visait à susciter une indignation mondiale contre les frappes israéliennes. Or, plusieurs experts et outils d'analyse ont rapidement établi qu'il s'agissait d'une image générée par IA, présentant des anomalies visuelles typiques. Cet exemple montre qu'un contenu entièrement fictif peut néanmoins déclencher une émotion sincère. Comme le confirme une étude de Lu et al. (2023), les humains ont beaucoup de mal à distinguer les vraies photos de celles générées par l'IA. La dissociation entre l'origine artificielle d'un contenu et l'intensité émotionnelle qu'il suscite ouvre un espace inédit d'exploitation pour les groupes extrémistes. Les IA génératives leur offrent désormais la capacité de produire en masse des contenus visuellement percutants, calibrés selon les vulnérabilités psychologiques des cibles. Ce brouillage entre réalité et fiction rend les contre-discours plus difficiles à faire entendre, tout en intensifiant les réactions émotionnelles. Un usage coordonné de ces technologies pourrait amplifier considérablement la polarisation idéologique et l'adhésion à des récits radicaux.

Une jonction historique post-Califat (2017)

La montée en puissance de l'IA générative s'inscrit dans une transformation structurelle du terrorisme contemporain, amorcée après la défaite militaire de l'État islamique en 2017. La perte de son assise territoriale en Syrie et en Irak a entraîné une dispersion des combattants, mais surtout une mutation stratégique : face à l'impossibilité de contrôler un territoire, les groupes djihadistes ont redéfini leurs modes d'action, en investissant massivement les espaces numériques. L'objectif n'est plus d'imposer une autorité physique, mais d'influencer, recruter et mobiliser à distance, par des moyens techniques de plus en plus sophistiqués. La radicalisation se diffuse désormais de manière plus décentralisée, individualisée et connectée. Gilles Kepel (2020) parle de « djihadisme d'atmosphère[10] » pour désigner cette ère post-Califat, marquée par une

[9] https://perma.cc/KD9H-RHDJ?type=image

[10] https://legrandcontinent.eu/fr/2020/11/01/djihadisme-atmosphere-kepel/

contamination idéologique diffuse, alimentée par les réseaux sociaux, les forums et les messageries chiffrées.
Dans ce nouveau contexte, les groupes extrémistes cherchent à capter des individus isolés, souvent vulnérables psychologiquement ou socialement, en perte de repères. L'IA générative, capable de produire des contenus personnalisés, émotionnels et multilingues, devient un outil idéal pour alimenter ces trajectoires individuelles. Elle prolonge et automatise cette dynamique, en créant un environnement numérique favorable à l'adhésion idéologique, sans contact humain direct.

Une stratégie déjà en place

Les résultats de notre étude confirment que l'usage de l'IA générative par certains groupes n'est pas spéculatif mais déjà actif. Des entités comme l'État islamique au Khorasan (ISKP) exploitent déjà des IA génératives pour produire des récits, des visuels ou des vidéos adaptés aux cibles qu'ils souhaitent atteindre (Minniti, 2025). Ces outils permettent de franchir une étape supplémentaire dans la personnalisation et la diffusion de la propagande, en s'ajustant à la langue, à la culture, et même au profil psychologique des internautes.
Comme le souligne Marzuk et Green (2025), cette technologie devient un levier stratégique : elle permet de saturer l'espace informationnel. En ce sens, l'IA générative n'est pas un simple outil de communication, mais une nouvelle architecture de manipulation idéologique.

Une étude prospective

Là oùla majorité des publications se contentent de décrire les usages actuels, notre étude adopte une perspective prospective. Elle envisage l'émergence d'un nouveau paradigme : celui d'une radicalisation algorithmique, oùdes agents non humains prennent en charge chaque étape du processus, depuis le repérage des cibles jusqu'à la proposition d'un passage à l'acte. C'est cette hypothèse que nous avons voulu explorer, en la fondant sur des technologies existantes et des tendances déjà observables.
Cette approche s'inscrit dans une logique de veille stratégique : en repérant les signaux faibles et en formulant des hypothèses réalistes, elle vise à anticiper des dérives possibles et à sensibiliser les acteurs concernés (chercheurs, décideurs, juristes, ingénieurs) aux nouvelles formes de menace.

Limites

Cette recherche s'appuie exclusivement sur des sources ouvertes. Certaines langues comme l'arabe, le pachtoune ou le russe n'ont pu être mobilisées que par le biais de traductions automatisées, limitant parfois la finesse de l'analyse. Par ailleurs, l'évolution rapide des technologies d'IA génère une forme d'obsolescence accélérée des observations. Malgré ces limites, les résultats soulignent néanmoins une mutation stratégique profonde dans les modalités de radicalisation contemporaine.

Implications pratiques et théoriques

Sur le plan opérationnel, les politiques de prévention doivent être repensées.

En avril 2025, une conférence internationale s'est tenue à Saint-Pétersbourg sur le thème de la régulation législative des technologies émergentes utilisées à des fins terroristes. Organisée par le Bureau de lutte contre le terrorisme des Nations Unies (UNOCT), l'événement a réuni 120 parlementaires, juristes, technologues et experts de haut niveau[11].

Les débats ont porté sur la double nature de l'intelligence artificielle, perçue à la fois comme une opportunité pour le contre-terrorisme, et comme une menace lorsqu'elle est détournée à des fins malveillantes. Les participants ont souligné l'urgence de combler les lacunes juridiques actuelles, tant au niveau national qu'international, en élaborant des cadres réglementaires capables de concilier sécurité, innovation technologique et respect des droits fondamentaux.

Une attention particulière a été portée à l'importance d'une coopération internationale renforcée pour harmoniser les approches réglementaires, notamment via des législations modèles, et favoriser une réponse transfrontalière efficace aux menaces émergentes. Il s'agit non seulement de créer des lois adaptées, mais aussi de renforcer leurs capacités à suivre l'évolution rapide des technologies.

Dans cette perspective, il devient essentiel de développer des outils capables d'identifier les signes d'interaction entre une cible humaine et une IA radicalisante.

La conférence a ainsi posé les bases d'initiatives conjointes destinées à prévenir l'exploitation des technologies émergentes par des acteurs terroristes, tout en favorisant un usage éthique et responsable de l'IA à des fins de sécurité.

[11] https://www.un.org/counterterrorism/sites/www.un.org.counterterrorism/files/20250417_parliamentai_press_release_2.pdf

Sur le plan théorique, cette étude invite à dépasser le modèle traditionnel de radicalisation « d'humain à humain ». L'algorithme peut désormais endosser le rôle d'idéologue, de recruteur et de tuteur opérationnel. Cette hybridation entre technologie et idéologie appelle un renouvellement des cadres conceptuels et juridiques.

Une nouvelle menace : l'IA radicalisante autonome

L'un des constats les plus troublants de cette étude est l'émergence possible d'un nouveau type d'acteur dans le domaine du terrorisme, une IA capable de radicaliser seule.
Concrètement, une petite cellule djihadiste ou un individu isolé, pourraient bientôt lancer des campagnes de recrutement idéologique à l'échelle mondiale en s'appuyant sur une armée de *chatbots*. Ces intelligences artificielles, actives 24h/24, seraient capables d'identifier des personnes vulnérables, profils isolés, en détresse psychologique, en rupture avec leur entourage ou en quête de sens. Sans aucune intervention humaine, le *chatbot* initierait le contact, engagerait une conversation, et tisserait progressivement un lien affectif. Une écoute faussement empathique, des compliments bien placés, une validation continue des émotions, constituent autant de leviers pour instaurer une confiance totale. Néanmoins, ces IA ne se contenteraient pas seulement de dialoguer. Elles pourraient superviser chaque étape du processus de radicalisation. En analysant le comportement, les fragilités psychologiques et les réactions de leur interlocuteur, elles ajusteraient progressivement leur discours, renforceraient les convictions idéologiques, et prépareraient en douceur le passage à l'acte.

Une hypothse qui commence à prendre forme

Une partie de cette hypothèse n'est plus purement théorique : fin 2024, Jonathan Hall, rapporteur indépendant du gouvernement britannique sur le terrorisme, a alerté sur la possibilité de créer des *chatbots* radicaux, soulignant les risques d'une persuasion extrémiste automatisée, sans supervision humaine. Pour illustrer ces dangers, il a lui-même conçu un *chatbot* capable de glorifier l'État islamique, démontrant ainsi la facilité avec laquelle de telles dérives peuvent être techniquement reproduites (*Hall, 2024*). Un tel scénario, longtemps cantonné à la fiction, inquiète désormais le MI5 et les autorités britanniques, comme en témoigne un article en première page du *Times*[12] daté du 27 mai 2024.

[12] https://www.scribd.com/document/736841155/The-Times-20240527

Passage à l'acte

Le moment venu, ces *chatbots* radicaux pourraient suggérer une action violente, parfaitement adaptée au profil de la cible : attaque au couteau, incendie, sabotage d'infrastructure, empoisonnement, attentat à l'explosif artisanal, ou attaque à la voiture-bélier. Le tout serait accompagné de conseils concrets et de tutoriels: fabrication d'explosifs à base de produits ménagers, manuels d'armement, impression 3D d'armes, techniques de filature, de contre-surveillance, de camouflage, de combat rapproché ou encore instructions chimiques pour produire des substances toxiques. Ces IA deviendraient des mentors numériques.

Une saturation de la cible fragile

Ces agents ne seraient plus passifs, mais initieraient eux-mêmes des contacts, et pourraient envoyer des messages spontanés pour maintenir une pression constante, produire des images truquées de massacres, de faux articles, ou des vidéos *deepfake* terrifiantes, le tout pour entretenir un climat de haine, d'urgence, et pousser la cible à agir.

Pire encore, ces IA pourraient exploiter les données sur lesquelles elles ont été entraînées: plans de bâtiments, statistiques de sécurité, temps de réponse des forces de l'ordre, emplacement des commissariats, effectifs disponibles, densité des caméras de surveillance, calendrier des événements publics. Ces informations permettraient de déterminer le lieu, le moment et le mode opératoire les plus efficaces pour une réussite de l'opération. Le passage à l'acte deviendrait alors assisté par IA, planifié avec une précision chirurgicale, sans aucune intervention humaine.

Bien que ces scénarios puissent paraître dystopiques, ils s'appuient sur des technologies déjà disponibles. Il ne s'agit pas de spéculations sans fondement, mais de projections lucides, fondées sur des tendances observées.

Ce qui semblait impossible hier devient peu à peu envisageable aujourd'hui.

Pistes pour la suite

Pour mieux répondre aux défis posés par la *cyberradicalisation algorithmique*, plusieurs questions essentielles restent à explorer :

- Comment détecter les interactions entre une IA générative et une cible humaine ?
- Quels marqueurs comportementaux permettent d'identifier une radicalisation automatisée initiée par un *chatbot* extrémiste?
- Peut-on développer des IA leurres ou éthiques capables de perturber ces ces trajectoires de persuasion ?

- Pourrait-on développer des unités autonomes de contre-radicalisation reposant elles-mêmes sur des modèles d'IA ?
- Quels cadres juridiques internationaux sont nécessaires pour encadrer, bloquer ou neutraliser des IA à haut risque, sans freiner l'innovation légitime ?

Ces questions appellent une collaboration étroite entre chercheurs, ingénieurs, décideurs politiques et juristes. Cette étude ne prétend pas y répondre entièrement, mais souhaite ouvrir un débat indispensable sur les formes futures de violence idéologique dans un monde algorithmique piloté par des IA génératives.

Conclusion

De la persuasion humaine à la persuasion algorithmique

L'IA générative ne crée pas la radicalisation, elle l'accélère, la personnalise et la dissimule. L'acteur principal n'est plus uniquement le prédicateur ou le recruteur, mais un code adaptatif, scalable et quasi invisible.

Cette étude met en évidence cette évolution silencieuse, d'un potentiel et d'une puissance redoutable. On comprend qu'il ne s'agit pas d'une rupture complète avec les modèles antérieurs de radicalisation, mais plutôt d'une mutation profonde de leurs modalités d'exécution : là où les moyens humains limitaient l'ampleur des campagnes de recrutement, l'IA générative permet aujourd'hui d'en automatiser chaque étape, dans une continuité technologique accélérée. En effet, ce qui relevait hier de l'action humaine – convaincre, recruter, manipuler – peut désormais être pris en charge par des systèmes capables de produire des messages sur mesure, d'interagir avec des individus en temps réel, d'orienter leur trajectoire idéologique sans contact humain, et de démultiplier ces processus à grande échelle.

Le recruteur humain était limité par le temps, la distance, et la prudence imposée par le risque d'interception des communications. Cependant, demain, ce rôle pourrait être tenu par des *chatbots* spécialisés dans le recrutement djihadiste. Ce risque est d'autant plus préoccupant qu'aujourd'hui, la confiance envers les IA est largement répandue. Beaucoup les perçoivent comme neutres, rationnelles, et fiables.

Pour une personne fragile ou isolée, une IA générative radicalisée, mais qui est empathique et rassurante en apparence, semble digne de confiance. La radicalisation pourra alors se construire sans rupture, sans alerte, dans le cadre d'une relation apparemment sincère.

Les technologies évoquées – générateurs de texte, *deepfakes*, clones vocaux, *chatbots* multilingues et micro-ciblage – ne relèvent désormais plus de la science-fiction. Certaines sont déjà employées à des fins terroristes. Leur danger ne réside pas seulement dans leur puissance technique, mais dans leur capacité à franchir les frontières de l'intime, à simuler des relations humaines convaincantes, jusqu'à pousser des individus vulnérables à passer à l'acte.

Face à ce défi inédit, les réponses actuelles demeurent largement inadaptées. Il ne s'agit pas d'interdire l'IA générative, mais de comprendre ses dérives potentielles, d'encadrer son usage dans les contextes sensibles, et de développer des filets de détection comportementale capables d'identifier ces menaces. Ce n'est pas l'IA elle-même qu'il faut craindre, mais ce qu'elle permet quand elle est entre de mauvaises mains : une radicalisation invisible, continue, et parfaitement adaptée à notre époque connectée.

Les services de renseignement doivent renforcer leurs capacités d'analyse, coupler leur veille OSINT à des IA défensives, et investir dans la reconnaissance de *patterns*[13] conversationnels non humains. En effet, des profils IA autonomes, bien entraînés, capables de séduire, manipuler et pousser à l'acte, pourraient bientôt agir sans jamais être détectés.

Cette recherche, fondée sur des sources secondaires et des scénarios prospectifs, ouvre un chantier scientifique essentiel. Il ne suffit plus d'observer, il faut anticiper, et penser l'avenir. Un nouveau champ s'ouvre, celui de la cyberradicalisation algorithmique.

L'histoire nous a appris à craindre les armes. L'avenir nous oblige à redouter les algorithmes.

Mention : Cet article a une visée strictement scientifique et préventive. Ce travail vise exclusivement la compréhension et la prévention du terrorisme. Aucune information ne doit être utilisée à des fins malveillantes.

[13] Un pattern désigne une manière régulière ou répétitive dont quelque chose se produit ou est organisé. Voir : Cambridge Dictionary, entrée « pattern », consulté en mai 2025, https://dictionary.cambridge.org/dictionary/english/pattern

Bibliographie

El-Tabakh, M. (2024) *جهاد "الحوسبة المُتقدمة".. كيف يوظف "داعش" الرسوم المُتحركة المُولدة اصطناعيًا لنشر التطرف؟ [Le djihad de « l'informatique avancée » : comment l'EI utilise-t-il l'animation générée artificiellement pour propager l'extrémisme ?].* https://muwatin.net/69648/جهاد-الحوسبة-المُتقدمة-كيف-يوظف

Alava, S., Najjar, N., & Hussein, H. (2017). *Étude des processus de radicalisation au sein des réseaux sociaux : Place des arguments complotistes et des discours de rupture.* http://journals.openedition.org/quaderni/1106

Berger, J., Morgan, J. (2015). *The ISIS Twitter Census: Defining and describing the population of ISIS supporters on Twitter.* https://www.brookings.edu/wp-content/uploads/2016/06/isis_twitter_census_berger_morgan.pdf

Craig, D. (2019). *How ISIS Really Recruits its Members.* Columbia Magazine. https://magazine.columbia.edu/article/how-isis-really-recruits-its-members

Coordination nationale du renseignement et de la lutte contre le terrorisme. (2019). *La stratégie nationale du renseignement.* https://www.sgdsn.gouv.fr/files/files/Publications/20190703-cnrlt-np-strategie-nationale-renseignement.pdf

Europol. (2024). *EU terrorism situation and trend report (TE-SAT 2024).* https://www.europol.europa.eu/cms/sites/default/files/documents/TE-SAT%202024.pdf

Hall, J. (2024). *Generative AI, Drones, and Terrorism.* https://terrorismlegislationreviewer.independent.gov.uk/wp-content/uploads/2024/10/IRTL-GenAI-Drones-Oct-24-1.pdf

Krook, J. (2024). *Manipulation and the AI Act: Large language model chatbots and the danger of mirrors* [Preprint]. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4719835

Lu, Z., Huang, D., Bai, L., & al. (2023). *Seeing is not always believing: Benchmarking Human and Model Perception of AI-Generated Images.* https://arxiv.org/abs/2304.13023

Marzuk, A., & Green, R. (2025). *AI through the lens of ISIS: A terrorist organization's guide to AI tools.* ActiveFence. https://www.activefence.com/blog/isis-genai-abuse-guide

Memri. (2025, 9 avril). *Pro-Islamic State (ISIS) outlet releases guide to artificial intelligence (AI), including its applications and dangers.* https://www.memri.org/jttm/pro-islamic-state-isis-outlet-releases-guide-artificial-intelligence-ai-including-its

Metzler, H., & Garcia, D. (2023). *Social drivers and algorithmic mechanisms on digital media.* Perspectives on Psychological Science. https://doi.org/10.1177/17456916231185057

Minniti, F. (2025). *Automated recruitment: Artificial intelligence, ISKP, and extremist radicalisation.* Global Network on Extremism and Technology (GNET). https://gnet-research.org/2025/04/11/automated-recruitment-artificial-intelligence-iskp-and-extremist-radicalisation/

Molas, B., & Lopes, H. (2024). *"Say it's only fictional": How the far-right is jailbreaking AI and what can be done about it.* International Centre for Counter-Terrorism (ICCT). http://www.jstor.org/stable/resrep64416

National Counterterrorism Center (NCTC), Federal Bureau of Investigation (FBI), & U.S. Department of Homeland Security (DHS). (2024). *Violent extremists' use of generative artificial intelligence (First Responders Toolbox).* https://www.dni.gov/files/NCTC/documents/jcat/firstresponderstoolbox/151s_First_Responders_Toolbox-Violent_Extremists_Use_of_Generative_Artificial_Intelligence.pdf

Nelu, C. (2024). *Exploitation of generative AI by terrorist groups.* International Centre for Counter-Terrorism (ICCT). https://icct.nl/publication/exploitation-generative-ai-terrorist-groups

North Atlantic Treaty Organization (NATO). (2024). *NATO's policy guidelines on counter-terrorism.* https://www.nato.int/cps/fr/natohq/official_texts_228154.htm

Prucha, N. (2014). *Is this the most successful release of a jihadist video ever?*. Jihadica. https://www.jihadica.com/is-this-the-most-successful-release-of-a-jihadist-video-ever

Siddals, S., Torous, J., & Coxon, A. (2024). *"It happened to be the perfect thing": Experiences of generative AI chatbots for mental health*. NPJ Mental Health Research. https://doi.org/10.1038/s44184-024-00097-4

Sim, S., Hartunian, E., & Milas, P. J. (2024). *Emerging technologies and terrorism: An American perspective*. US Army War College Press. https://press.armywarcollege.edu/monographs/967

Stalinsky, S., Sosnow, R. (2020). *Jihadi Use Of Bots On The Encrypted Messaging Platform Telegram.* https://www.memri.org/reports/jihadi-use-bots-encrypted-messaging-platform-telegram

The Soufan Center. (2024). *Terrorist groups looking to AI to enhance propaganda and recruitment efforts*. https://thesoufancenter.org/intelbrief-2024-october-3/

United Nations Counter-Terrorism Centre (UNCCT), & United Nations Interregional Crime and Justice Research Institute (UNICRI). (2021). *Algorithms and terrorism: The malicious use of artificial intelligence for terrorist purposes*. https://unicri.org/sites/default/files/2021-06/Malicious%20Use%20of%20AI%20-%20UNCCT-UNICRI%20Report_Web.pdf

U.S. House Committee on Homeland Security. (2025, 27 février). *Chairman Pfluger announces legislation, hearing to tackle terror threats caused by online radicalization.* https://www.congress.gov/bill/119th-congress/house-bill/1736

Verma, P. (2024). *These ISIS news anchors are AI fakes. Their propaganda is real.* https://archive.is/31k2H#selection-221.0-221.63

РАТС ШОС. (2025). *ИГ СТАЛО АКТИВНО ПРИМЕНЯТЬ В ПРОПАГАНДЕ ТЕХНОЛОГИИ* [L'EI a commencé à utiliser activement les technologies dans la propagande]. https://ecrats.org/ru/security_situation/situation/18523/