Global Network
on Extremism & Technology

Intelligence artificielle et lutte contre l'extrémisme violent : rapport introductif

Marie Schroeter

Le GNET est un projet spécial du Centre international d'étude de la radicalisation du King's College, à Londres.

L'auteure de ce rapport est Marie Schroeter, Mercator Fellow sur l'usage des nouvelles technologies dans les relations internationales : potentiels et limites de l'intelligence artificielle pour la prévention de l'extrémisme violent en ligne.

Le Global Network on Extremism and Technology (Réseau mondial sur l'extrémisme et la technologie – GNET) est une initiative de recherche universitaire bénéficiant du soutien du Forum mondial de l'Internet contre le terrorisme (GIFCT), une initiative indépendante mais financée par le secteur qui vise à mieux comprendre et lutter contre l'utilisation des technologies par les groupes terroristes. Le GNET est formé et dirigé par le Centre international d'étude de la radicalisation (ICSR), un centre de recherche universitaire basé dans les locaux du Département d'étude des guerres du King's College, à Londres. Les opinions et conclusions exprimées dans ce document sont celles des auteurs et ne doivent en aucun cas être interprétées comme représentant les opinions et conclusions, expresses ou implicites, du GIFCT, du GNET ou de l'ICSR.

Nous tenons à remercier Tech Against Terrorism pour le soutien apporté dans le cadre de ce rapport.

COORDONNÉES

Pour toute question, demande d'information et demande de copies supplémentaires du présent rapport, contacter :

ICSR
King's College London
Strand
Londres WC2R 2LS
Royaume-Uni

T. **+44 20 7848 2098**
E. **mail@gnet-research.org**

Twitter : **@GNET_research**

Ce rapport peut, comme toutes les autres publications du GNET, être téléchargé gratuitement à partir du site Internet du GNET : www.gnet-research.org.

Résumé exécutif

La radicalisation peut avoir lieu tant hors ligne que sur la toile. L'étendue du rôle joué par Internet reste contestée, mais il va sans dire qu'il existe des communautés radicalisées et extrémistes en ligne. Dans ce rapport, nous nous intéressons à la capacité de contribution des applications d'intelligence artificielle (IA) à la lutte contre la radicalisation. Nous recensons les possibilités offertes par ces technologies sous leurs différentes formes, ainsi que leurs limites, afin d'aider les décideurs et spécialistes à faire le tri entre les informations pertinentes et le bruit et à prendre ainsi des décisions éclairées et indépendantes vis-à-vis du battage médiatique actuel. Les conclusions listées ci-dessous ont été les plus marquantes :

1. La manipulation des moteurs de recherche et systèmes de recommandation peut contribuer à la lutte contre la radicalisation en dirigeant les internautes vers des contenus modérés

En réduisant les chances pour les internautes de tomber sur des contenus susceptibles d'entraîner une radicalisation, les moteurs de recherche et systèmes de recommandation peuvent contribuer à la prévention de l'extrémisme violent et ainsi jouer un rôle important dans la sécurisation des espaces en ligne. Les moteurs de recherche aident les utilisateurs à ne pas se noyer dans la masse d'informations en ligne, y compris les contenus à caractère extrémiste. L'utilisation d'algorithmes manipulés pourrait les diriger vers des contenus modérés plutôt qu'extrémistes. De même, les systèmes de recommandation qui suggèrent la prochaine vidéo, la prochaine chanson ou le prochain film à partir de l'historique de navigation sont susceptibles de renforcer les points de vue extrêmes en recommandant des contenus confirmatifs. Un système de recommandation équilibré permettrait de lutter contre les récits malveillants en proposant des contenus adverses, ou de diffuser des informations sur des projets ou des points de contact consacrés à la prévention et à la lutte contre l'extrémisme violent.

2. Le traitement automatique du langage naturel peut aider à traduire les langues minoritaires en vue d'améliorer la modération des contenus et permettre la modération des contenus de sites ultra spécialisés à long terme

Le traitement automatique du langage naturel (TALN) présente un potentiel pour la modération des contenus en ligne, notamment en ce qui concerne les langues parlées uniquement par de petits groupes de population. Souvent, la modération des contenus dans les langues minoritaires ne semble pas suffisamment rentable pour les investissements. Les plateformes plus modestes n'ont pas toujours l'expertise technique ou les ressources nécessaires pour se doter de systèmes de modération des contenus, puisque même

l'utilisation des modèles existants suppose beaucoup de temps et d'efforts. D'autres défendent ce qui pourrait être considéré comme une interprétation extrême de la valeur de la liberté d'expression et ne souhaitent pas restreindre leurs utilisateurs. Un système amélioré de TALN pourrait aider à traduire des contenus dans des langues utilisées par un grand nombre de modérateurs expérimentés et formés. Le TALN peut également détecter la présence de modèles sémantiques inhabituels en ligne. Cela pourrait favoriser le repérage de messages critiques sur les plateformes qui ne veulent ou ne peuvent pas investir dans la modération de contenu. Toutefois, ces mesures doivent garantir le respect permanent des normes de confidentialité et des droits humains.

3. La lutte contre la désinformation et les contenus manipulés en ligne manque de solutions automatisées

Il n'existe, à l'heure actuelle, aucun outil automatisé convaincant permettant d'identifier et de traiter la désinformation et les contenus manipulés néfastes mais licites. L'amélioration considérable de la culture numérique des utilisateurs, dans le but de créer une souveraineté numérique, semble être une approche plus utile à court terme.

4. L'IA surhumaine ne va pas « tirer la sonnette d'alarme » si un individu se radicalise en ligne

La mise en place d'une IA générale qui contrôle, à l'aide d'une intelligence surhumaine, les comportements et les contenus postés par les individus en ligne et « tire la sonnette d'alarme » en présence de plusieurs indicateurs d'une radicalisation n'est pas réaliste, et restera de l'ordre de la science-fiction, et ce pour deux raisons. Tout d'abord, il n'existe pas suffisamment de données pour alimenter un algorithme avec des informations précises sur la radicalisation et sur le moment où un individu radicalisé se tourne vers la violence. À moins qu'une innovation technique permettant la création de systèmes fiables fondés sur un nombre de données plus restreint ne soit introduite, il n'existe aucune incitation à utiliser l'assistance technique, puisqu'elle ne serait pas capable de faire des prédictions fiables en l'absence de données suffisantes sur les cas antérieurs. Les cas de radicalisation et de terrorisme sont – heureusement – trop rares et variés pour produire suffisamment d'informations pour un algorithme. Deuxièmement, la prédiction du comportement des individus nécessiterait d'avoir des données clairement attribuables sur les individus, ce qui trahirait tous les aspects de la confidentialité et pourrait résulter en une surveillance d'une ampleur sans précédent. Le scénario décrit ci-dessus n'est pas compatible avec les démocraties libérales, dont les fondations reposent sur le droit à la vie privée.

Table des matières

1 Introduction

Selon les croyances populaires, l'intelligence artificielle (IA) va tout révolutionner, y compris la sécurité nationale. La mesure dans laquelle Internet favorise la radicalisation demeure une question sans réponse, mais les dernières attaques terroristes, à Halle dans l'est de l'Allemagne, à Christchurch en Nouvelle-Zélande ou encore dans la synagogue de Poway, en Californie, ne sont que trois illustrations récentes du rôle joué par la sphère numérique dans la radicalisation aujourd'hui.

Comment l'IA peut-elle aider à lutter contre la radicalisation en ligne ? L'expertise en la matière ne se cantonne pas à une seule discipline ; on retrouve ainsi des spécialistes parmi les chercheurs et experts issus des secteurs de la sécurité et de la lutte contre le terrorisme, ainsi que parmi les décideurs politiques et experts techniques, qui unissent leurs forces de plus en plus pour explorer ce domaine. Le paysage actuel d'informations ne facilite pas la tâche des décideurs politiques en matière de filtrage des informations. L'objectif de ce rapport est de faire la lumière sur les dernières évolutions en matière d'IA et de les placer dans le contexte de la lutte contre la radicalisation dans les démocraties libérales.

Cette publication contribue au sujet en soulignant certaines des limites et possibilités de l'IA en matière de lutte contre la radicalisation en ligne. Le deuxième chapitre explique brièvement les principaux concepts et les principales idées sous-tendant l'IA. Dans la section « Pour aller plus loin » à la fin du chapitre, nous concentrons notre attention sur la qualité des données, ainsi que sur les problèmes de subjectivité et la manipulation dans les ensembles de données. Le troisième chapitre traite du potentiel offert par les innovations technologiques fondées sur l'IA – ainsi que de leurs limites – pour la garantie d'un espace en ligne « sain », exempt de contenu à caractère terroriste, de propagande et de faux engagements. L'hypothèse est que cet environnement numérique sain contribue à la prévention de la radicalisation. Ce chapitre évalue un ensemble de concepts populaires fondés sur l'IA, allant des vidéos manipulées (deepfakes) aux armées de robots diffusant de fausses informations, et explique pourquoi les moteurs de recherche, les systèmes de recommandation et, en particulier, le traitement automatique du langage naturel (TALN) ont le potentiel de contribuer à cet objectif d'une façon ou d'une autre. Le quatrième chapitre s'intéresse uniquement à une « IA générale » hypothétique, un système omniscient qui identifierait les individus en cours de radicalisation et aiderait par conséquent les autorités à prévenir la commission d'infractions. Ce chapitre soutient également qu'une telle technologie ne peut toutefois relever que de la science-fiction, du moins pour les années à venir. Ce point débouche sur une discussion portant sur les raisons sous-tendant cette prise de position. Les débats sur les mégadonnées, notamment en ce qui concerne la sécurité traditionnelle, ne peuvent avoir lieu dans les démocraties libérales sans assurer la protection de la vie privée et lui donner la priorité. Un autre segment « Pour aller plus loin » dans le quatrième chapitre fournit des informations plus approfondies pour toute personne intéressée. Le cinquième chapitre conclue le rapport.

Le présent rapport est fondé sur des entretiens semi-structurés avec des chercheurs, des décideurs et conseillers politiques, ainsi qu'avec des représentants du secteur privé. Par ailleurs, les conclusions de recherches documentaires et d'exercices de suivi des médias ont influencé les prises de positions communiquées ici. J'ai parlé à différentes parties prenantes pour obtenir une perspective pluridisciplinaire tenant compte du paysage informatif fragmenté. Toutefois, cette recherche présente des limites évidentes, résultant du fait que les informations sur l'utilisation de l'apprentissage automatique ont été soit restreintes par les services de renseignement, soit limitées par les entreprises du secteur privé.

2 Qu'est-ce que l'intelligence artificielle ?

L'intelligence artificielle (IA) a beau être un terme très à la mode aujourd'hui, il n'existe pourtant pas de définition standard universelle. Ceci est en partie dû au fait que l'étude de l'IA est un sujet très populaire qui évolue à toute vitesse, qui produit constamment de nouvelles conclusions et brouille les frontières entre informatique, statistique et robotique. Bien qu'il n'existe pas de consensus sur sa définition, l'intelligence artificielle touche la plupart d'entre nous. C'est ce type de technologie qui nous recommande quels biens acheter en ligne, qui gère notre planning et qui conduit les véhicules autonomes.

Alexa, l'assistant vocal d'Amazon, nous montre le degré de complexité que peuvent atteindre les systèmes de prise de décision automatisée : Alexa peut, de façon très pratique, prévoir une soirée romantique de A à Z, en réservant des billets pour un spectacle et une table au restaurant, en commandant un Uber et en informant votre partenaire de votre heure d'arrivée[1]. Plus généralement, le terme IA décrit une discipline qui traite de systèmes technologiques automatisés et évolutifs. L'IA effectue des tâches en l'absence de contrôle permanent et est capable d'améliorer ses performances en tirant des leçons des expériences antérieures[2].

C'est lors d'une conférence au Dartmouth College, dans la ville de Hanover, dans le New Hampshire (États-Unis), en 1956, que ces technologies ont été baptisées « intelligence artificielle ». Après quelques succès initiaux, les chercheurs se sont montrés optimistes quant aux progrès rapides pouvant caractériser l'utilisation d'algorithmes informatiques. Dès le début, ils ont pu écrire du code capable de résoudre des problèmes ; les programmes ainsi créés contenaient des éléments qui amélioraient les performances grâce à l'apprentissage. Compte tenu toutefois du manque de capacité de la mémoire et des processeurs, une « période noire » de l'IA s'est ensuivie, avec le gel des investissements et la perte d'intérêt dans ces recherches dans les années 1960.

Le battage médiatique dont jouit l'IA au 21e siècle s'explique par les évolutions techniques principalement pilotées par le secteur privé. Des solutions de stockage de masse et de logiciels de moins en moins coûteuses, associées au recrutement d'experts en la matière et d'un meilleur accès aux données, ont permis au secteur de prospérer. Mais qu'a l'IA de si spécial ? Tout d'abord, elle facilite l'analyse de données de masse, qui est plus rapide et plus efficace que celle effectuée par des opérateurs humains. Ensuite, elle est capable de travailler malgré les incertitudes et, grâce à cela, de prédire l'avenir. Le fait que ces prédictions soient ou non fiables est

1 Hao, K. (2019a), « Inside Amazon's plan for Alexa to run your entire life », MIT Technology Review. Disponible à l'adresse : https://www.technologyreview.com/s/614676/amazon-alexa-will-run-your-life-data-privacy/
2 Reaktor et Université d'Helsinki (2018), « How should we define AI? ». Disponible à l'adresse : https://course.elementsofai.com/1/1

secondaire. C'est précisément leur capacité à prédire l'avenir qui fait la force des algorithmes. Par comparaison, le cerveau humain est incapable de prendre des décisions fondées sur de vastes ensembles de données, des conditions multiples et des incertitudes. Le pouvoir de prédiction peut être considéré comme une capacité propre aux algorithmes.

Mettre l'accent sur le terme d'« intelligence artificielle » est, en soi, trompeur. Cela suggère en effet qu'il existe des similarités avec l'intelligence humaine ou les processus d'apprentissage humains. Les systèmes neuronaux profonds, une technique spéciale d'apprentissage automatique présentant plusieurs couches d'unités de traitement, sont certes inspirés par l'architecture du cerveau humain. Leurs capacités sont toutefois très différentes de celles des neurones humains. Même dans les cas complexes, l'algorithme est capable de compléter les données manquantes à l'aide de modèles de prédiction, sans toutefois pouvoir donner un sens à ses conclusions. Les différences entre intelligence humaine et artificielle deviennent évidentes lorsque l'on réfléchit à ce qu'un réseau neuronal peut et ne peut pas faire. Par exemple, un algorithme peut déceler un cancer du sein au stade précoce de façon plus fiable que les humains, en analysant les images de mammographies avec une marge d'erreur plus faible que les radiologues[3]. En revanche, il ne peut pas comprendre les émotions de la patiente – c'est-à-dire y donner un sens – et réagir en conséquence. L'empathie nécessite des années d'observation et suppose de faire preuve d'intelligence émotionnelle, ce qui fait défaut dans les algorithmes. De plus, le mot « intelligence » utilisé dans le terme « IA » suppose que le système est capable de produire des pensées originales, ce qui est évidemment absurde. Le programme d'IA AlphaGo de Google peut facilement calculer les coups les plus prometteurs dans le jeu très complexe de go, mais cela ne veut pas dire qu'il comprend le jeu[4]. AlphaGo est incapable d'expliquer le contexte de ces coups, le fait qu'il joue à un jeu ou les raisons mêmes pour lesquelles on voudrait jouer à des jeux. L'attribution d'un sens et d'un contexte est une caractéristique très humaine ; la plupart des enfants sont capables de dire pourquoi c'est amusant de jouer à des jeux. Si le système est incapable d'expliquer pourquoi il fait ce qu'il fait, il est toutefois en mesure d'identifier les meilleurs coups dans une situation donnée en fonction de l'objectif que son programme doit atteindre. Il analyse toutes les options et décide mathématiquement comment minimiser le risque et donc faire face à l'incertitude.

La mise en lumière des forces et faiblesses de l'IA permet de découvrir l'existence de deux catégories. AlphaGo peut être considéré comme un système d'IA « étroit » ou « faible » puisqu'il ne gère qu'une seule tâche, alors qu'une IA « générale » serait capable de traiter n'importe quelle tâche de nature intellectuelle (ce type d'IA n'existe actuellement que dans la science-fiction). L'intelligence d'un système peut être faible ou forte, ce qui se traduit respectivement par une IA étroite ou générale. Le terme « IA étroite » (ou faible) renvoie à un système qui fait semblant d'être intelligent en produisant les résultats souhaités. L'intelligence peut être superficielle et souvent

3 Hao, K. (2020), « Google's AI breast cancer screening tool is learning to generalize across countries », MIT Technology Review. Disponible à l'adresse : https://www.technologyreview.com/f/615004/googles-ai-breast-cancer-screening-tool-is-learning-to-generalize-across-countries/

4 Gibney, E. (2017), « Self-taught AI is best yet at strategy game Go », Nature. Disponible à l'adresse : https://www.nature.com/news/self-taught-ai-is-best-yet-at-strategy-game-go-1.22858

fondée sur de fausses bases : par exemple, les algorithmes entraînés pour identifier des trains sur des photos ne sont en réalité pas capables de désigner un train, mais reconnaissent plutôt les voies ferrées parallèles souvent présentes dans les photos de trains. L'algorithme a pu se reposer sur de fausses structures dans ses réseaux neuronaux parce qu'elles ont produit le résultat souhaité[5]. Cela présente évidemment un risque : nous ne sommes pas encore totalement conscients des conséquences imprévues que cela pourrait engendrer. L'une des conséquences connues, par exemple, est le fait que les systèmes de reconnaissance faciale ont des difficultés à reconnaître les personnes noires[6]. Une IA générale aurait un véritable esprit, une conscience ou une intelligence supérieure, à laquelle pourraient se référer les médias populaires. Je le répète, les systèmes superintelligents n'existent, à l'heure actuelle, que dans la science-fiction.

La signification du terme IA a changé au fil du temps. Aujourd'hui, IA et apprentissage automatique (AA) sont utilisés indifféremment par de nombreux médias ; c'est également le cas de ce rapport. Il existe généralement deux domaines populaires d'AA : supervisé et non supervisé. L'AA supervisé signifie que l'algorithme est entraîné à analyser des données à partir d'un ensemble de données d'entraînement composé de données précédemment étiquetées. L'étiquetage peut s'entendre comme le choix entre « les données remplissent les conditions » et « les données ne remplissent pas les conditions ». Les données étiquetées donnent par conséquent un sens aux points de données, qui nécessitent une contribution humaine. Prenons, par exemple, l'image d'une pomme. Une telle image serait étiquetée « pomme ». Pour entraîner un algorithme, il est nécessaire d'avoir accès à une grande quantité de données clairement étiquetées. La performance de l'algorithme peut être évaluée à l'aide d'un ensemble de données test ; l'algorithme pourra alors, si l'expérience est jugée réussie, appliquer des étiquettes à de nouvelles données. L'avantage de l'apprentissage supervisé est que l'algorithme organise les données exactement comme cela a été programmé. L'étiquetage manuel des données est une tâche intense et coûteuse. De nombreux internautes étiquettent des données eux-mêmes, sur les sites qui posent des questions de sécurité visant à déterminer que « je ne suis pas un robot ». Ces questions peuvent, par exemple, demander à l'internaute de cocher toutes les images, sur un ensemble de photos, montrant une voiture. À cet instant précis, l'internaute étiquette des données. Le système reCaptcha de Google, qui propose un tel service de validation, utilise ces résultats pour entraîner les ensembles de données utilisés aux fins d'AA[7]. Par exemple, les données étiquetées pourraient servir à alimenter l'IA conduisant des véhicules autonomes.

En ce qui concerne l'AA non supervisé, l'algorithme doit repérer des tendances dans des données non étiquetées et non organisées sans avoir été entraîné à comprendre comment les données d'entrée correspondent aux données de sortie. Les algorithmes non supervisés repèrent des tendances au sein d'ensembles de données en l'absence d'autres consignes ou classifications. La difficulté réside

5 Thesing, L. *et al.* (2019), « What do AI algorithms actually learn? – On false structures in deep learning », Arxiv. Disponible à l'adresse : https://arxiv.org/abs/1906.01478
6 Simonite, T. (2019), « The best Algorithms Struggle to Recognize Black Faces Equally », Wired. Disponible à l'adresse : https://www.wired.com/story/best-algorithms-struggle-recognize-black-faces-equally/
7 Google ReCaptcha (n.d.). Disponible à l'adresse : https://www.google.com/recaptcha/intro/v3.html

dans le fait qu'il n'est pas possible de déterminer clairement si les tendances repérées seront utiles aux programmeurs de l'algorithme. Les conclusions risquent de ne pas être du tout pertinentes. Cette méthode permet tout de même d'éviter un étiquetage manuel intense des données et emploie la grande quantité de données non étiquetées ouvertement disponibles. L'apprentissage non supervisé peut servir comme première étape avant de travailler avec un algorithme d'apprentissage supervisé.

Il est souvent question d'apprentissage profond dans les reportages et informations. Il s'agit d'une technique spéciale d'AA, dans le cadre de laquelle plusieurs unités de traitement sont connectées en réseau ; l'échelle du réseau permet l'analyse de problèmes plus complexes. Les réseaux neuronaux font eux aussi souvent la une. Ceux-ci sont inspirés de la structure du cerveau humain et permettent de stocker et traiter les informations de façon simultanée. Ces réseaux ont joué un rôle important dans la percée du « big data », puisqu'ils ont permis de traiter de grandes quantités de données.

Pour aller plus loin

Données – qualité, subjectivité et manipulation

Il est rare de ne pas entendre la phrase d'accroche suivante dans les réunions ou conférences consacrées à l'IA : « les données numériques sont le nouveau pétrole ». Mais est-ce vraiment le cas ? Certes, l'AA est gourmand en données, et les données coûtent cher. De même, plus les algorithmes s'améliorent, plus ils ont de données. Toutefois, tous les points de données n'ont pas la même valeur, puisque la loi des rendements décroissants s'applique. Un algorithme en apprend davantage des premières données qui l'alimentent que de la millionième répétition. De même, les algorithmes travaillent particulièrement bien si les données d'entraînement les ont préparés à de nombreuses éventualités – c'est-à-dire, s'ils ont été alimentés par un ensemble de données qui comprend à la fois des éléments habituels et inhabituels. Les données sont donc le nouveau pétrole si l'on tient compte des prix élevés, de la forte demande (nombreux sont les nouveaux modèles commerciaux qui dépendent des données, puisqu'ils sont fondés sur l'AA) et, à l'heure actuelle, du petit nombre de détenteurs de la marchandise. Toutefois, si chaque goutte de pétrole contribue à la production pétrolière totale, les différents points de données exercent, quant à eux, une influence variable sur la valeur totale des données.

Les algorithmes sont vulnérables à la subjectivité, une déformation systématique de la réalité due à l'emploi d'échantillons de données. Un parti pris à l'entrée entraînera inévitablement un parti pris à la sortie et produira des effets multiplicateurs s'il est renforcé. L'existence d'une subjectivité – préjudiciable notamment pour les femmes et les noirs – dans les ensembles de données actuels a été mise en évidence à de nombreuses reprises. Par exemple, Amazon a dû retirer son outil automatisé de gestion des ressources humaines, qui s'avérait discriminatoire envers les femmes. L'algorithme attribuait en effet de meilleurs notes aux hommes qu'aux femmes car il avait été alimenté par les candidatures reçues lors des dix années précédentes. Ce n'est un secret pour personne que le secteur des nouvelles technologies est largement dominé par les hommes, un parti pris qui a été reproduit dans les décisions de l'algorithme[8].

Il est essentiel d'alimenter les algorithmes avec des données de qualité et non biaisées. À la suite d'un processus consultatif, les Nations Unies ont fondé l'organisation Tech Against Terrorism, qui a mis en place la Plateforme d'analyse des contenus à caractère terroriste (TCAP) visant à créer un ensemble de données répondant à ces critères et consultable par le secteur privé, les chercheurs et la société civile.
À l'origine, Tech Against Terrorism avait pour seul but d'inclure

8 Dastin, J. (2018), « Amazon scraps secret AI recruiting tool that showed bias against women », Reuters Technology News. Disponible à l'adresse : https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

des contenus d'Al-Qaïda et de l'État islamique (EI), mais a ultérieurement annoncé qu'elle étendrait la Plateforme pour y inclure également le terrorisme d'extrême droite. D'autres domaines, tels que le terrorisme alimenté par une idéologie misogyne – comme en témoigne l'existence du mouvement Incel – méritent tout autant d'attention. Évidemment, ces ensembles de données doivent respecter les normes de sécurité des données et prendre en compte les éventuelles conséquences pour la santé mentale des évaluateurs.

Cependant, les données peuvent être manipulées. Un ensemble de données qui a été falsifié est difficile à détecter de l'extérieur, notamment par les personnes manquant d'expertise technique. Les données n'ont pas nécessairement besoin d'être beaucoup modifiées pour manipuler un algorithme, comme l'ont prouvé des chercheurs chinois. Leurs expériences ont poussé un véhicule autonome à conduire sur le mauvais côté de la route. Ceci illustre bien la vulnérabilité des systèmes[9]. Pour les applications d'IA conçues pour prévenir la radicalisation en ligne, la manipulation des données pourrait entraîner le résultat inverse. On peut imaginer, par exemple, un système de modération de contenu floué qui décide de retirer les profils du parti politique d'opposition pendant la période précédant une élection. Une autre faiblesse découle de l'entraînement concurrentiel des systèmes d'AA, dans le cadre duquel deux systèmes se concurrencent, améliorant ainsi leurs performances réciproques. Par exemple, un système d'AA crée de faux visages tandis que l'autre doit les filtrer et les retirer d'un ensemble de vrais visages. À mesure que le système de filtrage s'améliore, le système de création de faux visages développe lui aussi ses compétences. Les conséquences de ce phénomène ne sont pas encore bien connues.

9 Knight, W. (2019), « Military artificial intelligence can be easily and dangerously fooled », MIT Technology Review. Disponible à l'adresse : https://www.technologyreview.com/2019/10/21/132277/military-artificial-intelligence-can-be-easily-and-dangerously-fooled/

3 IA et lutte contre la radicalisation en ligne – une question de détails

Comment les statistiques et la prise de décisions automatisée contribuent-elles à la lutte contre la radicalisation en ligne ? Comme toujours, tout est une question de détails, et le bruit autour de l'IA est assourdissant. Il est par conséquent difficile, pour les personnes non spécialisées, d'en comprendre l'ampleur réelle. Ce chapitre examine de près les possibilités et limites des innovations technologiques populaires fondées sur l'AA. Il porte sur les éléments qui dominent le débat public. Des vidéos manipulées (deepfakes) à la modération automatisée de contenu, des moteurs de recherche au traitement automatique du langage naturel, ce chapitre aide à évaluer le rôle joué par les technologies dans la lutte contre la radicalisation. Les chevauchements entre différents éléments sont inévitables dans un environnement aussi changeant qui connaît de nouveaux développements aussi fréquents.

3.1 Modeler l'expérience des internautes – ce qui est visible et facile à trouver

L'apprentissage automatique peut avoir une influence importante sur l'expérience des internautes et modeler ce qu'ils voient et trouvent facilement. Les algorithmes, quelle que soit la forme qu'ils prennent, présentent un potentiel important pour lutter contre la radicalisation en contribuant à la mise en place d'un espace virtuel plus sain et en prévenant la publication de contenus malveillants. Le présent rapport s'intéresse aux moteurs de recherche, aux systèmes de recommandation et à la modération automatisée de contenu.

Les **moteurs de recherche** aident les internautes à s'y retrouver parmi les millions de sites existants et à trouver des contenus pertinents en ligne. Tout comme un annuaire du 21e siècle, ils tracent la voie dans la masse d'informations et de données existant sur la toile. Les algorithmes derrière les moteurs de recherche sont la clé de leur succès. Il y a dix ans, le panorama des moteurs de recherche était beaucoup plus diversifié, mais la recette secrète de Google – développer la confiance des internautes en ses outils en leur fournissant des contenus pertinents – a finalement propulsé l'entreprise à la tête du secteur. Aujourd'hui, le moteur de recherche répond à des milliards de requêtes tous les jours, dont 15 % de nouvelles demandes. Non seulement ses algorithmes faciles à utiliser trouvent l'information recherchée, mais reconnaissent également les fautes d'orthographe et proposent automatiquement le prochain mot dans la barre de recherche. Au bout du compte, la façon dont l'algorithme est programmé décide des informations à présenter. L'accessibilité en ligne de manuels de fabrication de bombes artisanales aurait prétendument directement provoqué la commission d'activités terroristes, comme le montre le cas de Noelle Velentzas et Asia Siddiqui qui, d'après un agent du FBI, se sont servies du

magazine *Inspire* d'Al-Qaïda, d'articles de blog sur les explosifs artisanaux et de l'ouvrage *The Anarchist Cookbook* pour fabriquer des explosifs artisanaux[10]. La Grande-Bretagne a mis en place l'opération Cupcake, dans le cadre de laquelle les services de renseignements extérieurs (MI-6) et électroniques (GCHQ) ont remplacé un guide de fabrication de bombes artisanales dans le magazine *Inspire* par les recettes des meilleurs cupcakes américains. En résumé : la possibilité de trouver des manuels de fabrication de bombes artisanales ou la manipulation des algorithmes pour compliquer la découverte de contenus extrémistes en ligne revêtent une importance capitale. Cela n'arrêtera certes pas complètement les internautes ayant de solides compétences dans le domaine technologique, mais relève toutefois le niveau des obstacles à l'accès à de tels contenus.

Les **systèmes de recommandation** sont un outil pratique pour trouver la prochaine vidéo à visionner, la prochaine chanson à écouter, le prochain article de presse à lire ou le prochain achat à faire à partir des éléments consommés ou acquis par le passé. Ils peuvent mener à la découverte d'une nouvelle chanson ou faciliter la recherche de la literie adaptée à notre matelas tout neuf. Les algorithmes qui suggèrent les prochains éléments à consulter ou à écouter peuvent cependant aussi créer des **bulles de filtres** susceptibles de renforcer les croyances en proposant des contenus analogues. Ils peuvent par conséquent aussi renforcer les attitudes extrémistes. L'opacité avec laquelle les algorithmes suggèrent de nouveaux éléments sur les réseaux sociaux et les sites de streaming musical ou vidéo ne donne pas au consommateur la possibilité d'influencer la manière dont les choses peuvent lui être recommandées : plus de contenu du même type, perspectives opposées ou une autre combinaison des deux. Comme l'a montré une étude de 2019 sur les contenus autosuggérés des grosses plateformes de médias sociaux, les algorithmes de YouTube ont particulièrement contribué au renforcement des points de vue extrémistes. En effet, une fois qu'une vidéo contenant des contenus extrémistes ou à la marge a été visionnée, la plateforme recommande de visionner des contenus similaires[11]. C'est particulièrement alarmant, puisque YouTube est le média social le plus utilisé par la population adulte des États-Unis, et il n'est pas improbable que de nombreux utilisateurs s'informent à travers ce site[12].

Les médias sociaux grand public ont régulièrement été critiqués pour leur manque d'actions fermes contre l'exploitation de leurs plateformes par les terroristes. Il a été affirmé que les intermédiaires devaient assumer la responsabilité de la gestion de ces contenus, compte tenu du rôle d'espace quasi-public joué par les médias sociaux, espace dans lequel les gens se rencontrent, échangent des arguments et mènent des activités commerciales. Toutefois, d'innombrables **services de niche** offrent aujourd'hui un éventail diversifié de services en ligne sur l'ensemble du spectre de l'infrastructure numérique, allant des messageries garantissant un niveau élevé d'anonymat ou des plateformes de niche n'ayant ni les capacités, ni la volonté de contrôler le contenu aux services d'hébergement permettant la diffusion de

10 Tribunal de district des États-Unis, District Est de New York (2015), États-Unis d'Amérique c/ Noelle Velentzas et Asia Siddiqui. Plainte et déclaration sous serment à l'appui du mandat d'arrêt, 2014R00196. Disponible à l'adresse : https://www.justice.gov/sites/default/files/opa/press-releases/attachments/2015/04/02/velentzas-siddiqui-complaint.pdf

11 Reed *et al.* (2019), « Radical Filter Bubbles », dans : The 2019 GRNTT Series, an Anthology, RUSI, Londres.

12 Perrin, A. et Anderson, M. (2019), « Share of U.S. adults using social media, including Facebook, is mostly unchanged since 2018 », Pew Research Centre. Disponible à l'adresse : https://www.pewresearch.org/fact-tank/2019/04/10/share-of-u-s-adults-using-social-media-including-facebook-is-mostly-unchanged-since-2018/

manifestes et de vidéos en direct. Les sites Internet et services liés au secteur du jeu en ligne ont récemment essuyé de nombreuses critiques pour n'avoir pas empêché l'usage malveillant de leurs plateformes[13]. L'Université de Swansea, au Royaume-Uni, a montré dans une étude récente comment le réseau de l'EI avait exploité différents services d'hébergement pour son journal en ligne Rumiyah, décentralisant ainsi le contenu et augmentant la difficulté pour le retirer de façon efficace et rapide[14]. Il est urgent de mener davantage de recherches et d'obtenir des métadonnées sur l'utilisation des services de niche par les groupes terroristes.

3.2 Gestion des contenus créés par l'utilisateur

Le web 2.0 a révolutionné les interactions en ligne. Il a permis à un grand nombre d'internautes du monde entier de délaisser les sites Internet statiques au profit des interactions en temps réel. Si cette connectivité mondiale d'une ampleur sans précédent a renforcé et soutenu de nombreuses communautés, elle a également créé de nouvelles difficultés pour la lutte contre la radicalisation.

La **modération automatisée de contenu** sur les médias sociaux vise à prévenir la diffusion de contenu à caractère terroriste. À l'heure actuelle, 98 % des contenus malveillants postés sur Facebook sont déjà filtrés par les algorithmes d'AA, comme l'indique le dernier rapport d'auto-évaluation de l'entreprise destiné à l'UE sur la pratique de la désinformation[15]. Les utilisateurs signalent les 2 % restants. Twitter déclare remettre en cause dix comptes par seconde[16] ; Google, qui détient YouTube, supprime 80 % des vidéos inappropriées avant qu'elles ne soient visionnées, d'après ses propres informations[17]. En apparence, tout ceci semble constituer un modèle de modération réussi ; il est donc juste d'affirmer que la modération de contenu s'est améliorée ces dernières années. Il reste toutefois des failles énormes, en particulier lorsque l'on quitte les zones linguistiques standards. À l'heure actuelle, seules 14 des 26 langues officielles de l'Europe sont couvertes par le répertoire linguistique de vérification des faits de Facebook. Grâce à la sous-traitance, 15 pays africains sont aujourd'hui suivis, chiffre qui représente toutefois toujours moins d'un tiers des pays du continent[18] et qui ne donne aucune indication permettant de savoir s'il concerne uniquement les langues officielles ou inclut également les dialectes. L'absence de modération des contenus publiés dans les langues parlées par des minorités est un phénomène connu dans d'autres régions et pays très peuplés et diversifiés, comme l'Inde[19].

13 Schlegel, L. (2020), « Points, Ratings and Raiding the Sorcerer's Dungeon: Top-Down and Bottom-Up Gamification of Radicalisation and Extremist Violence ». Disponible à l'adresse : https://gnet-research.org/2020/02/17/points-rankings-raiding-the-sorcerers-dungeon-top-down-and-bottom-up-gamification-of-radicalization-and-extremist-violence/

14 Macdonald, S. *et al.* (2019), « Daesh, Twitter and the Social Media Ecosystem », The RUSI Journal, vol. 164, n° 4, p. 60-72.

15 Facebook (2019), *Facebook report on the implementation of the Code of Practice for Disinformation*. Disponible à l'adresse : https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=62681

16 Twitter (2019), *Twitter Progress Report: Code of Practice against Disinformation*. Disponible à l'adresse : https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=62682

17 Google (2019), *EC EU Code of Practice on Disinformation*. Disponible à l'adresse : https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=62680

18 Africa Times (2019), « Facebook expands fact-checking to 15 African nations ». Disponible à l'adresse : https://africatimes.com/2019/10/10/facebook-expands-fact-checking-to-15-african-nations/

19 Perrigo, B. (2019), « Facebook Says It's Removing More Hate Speech Than Ever Before. But There's a Catch », *Time*. Disponible à l'adresse : https://time.com/5739688/facebook-hate-speech-languages/

Comme nous l'avons vu précédemment, les algorithmes ne peuvent donner un sens aux données. Cela signifie qu'ils ne comprennent pas le contexte dans lequel prennent vie les comportements malveillants. Des exemples du Sri Lanka illustrent l'incapacité des algorithmes de Facebook à évaluer un contexte culturel à plusieurs niveaux. Avant les attentats à la bombe qui ont endeuillé Colombo à Pâques 2019, certaines publications sont passées au travers des mailles du filet parce que les algorithmes avaient été incapables de comprendre la complexité du discours haineux qu'elles contenaient. Même après plusieurs efforts pour signaler les propos haineux tenus, qui alimentaient un sentiment antimusulman polarisé, Facebook n'a pas réussi à supprimer le contenu ou à y répondre à l'aide d'une modération de contenu adaptée[20]. Pour classer l'argot utilisé comme discours de haine, n'importe quel algorithme chargé de modérer du contenu aurait dû être capable de comprendre les origines ethniques des parties concernées. Le problème va plus loin : souvent, les langues qui n'utilisent pas l'alphabet latin sont « traduites » dans cet alphabet à des fins pratiques. Certaines langues n'ont pas de futur grammaticalement explicite. À quoi ressemble donc une menace, qui, par définition, implique l'avenir ? Il conviendra de tenir compte de ces défauts de conception pour renforcer le filtrage automatisé.

De nombreuses entreprises de réseaux sociaux souhaitent empêcher que leurs plateformes soient exploitées par des acteurs malveillants. L'efficacité des algorithmes dans la détection de la propagande ou des activités terroristes dépend aussi de la qualité et de la disponibilité des données avec lesquelles ils ont été entraînés. Facebook a reconnu que l'absence de données d'entraînement expliquait pourquoi son algorithme n'avait pas réussi à repérer et filtrer les diffusions en direct de fusillades comme celle de Christchurch. L'entreprise utilise désormais les images des caméras portées par les fonctionnaires de la police britannique lors de leurs exercices d'entraînement à la lutte antiterroriste[21].

Une autre innovation fondée sur l'AA qui pourrait aider à gérer les contenus créés par les utilisateurs et améliorer la modération des contenus est le **traitement automatique du langage naturel** (TALN). Ceci décrit les procédures et outils techniques mis en place pour analyser et traiter les langues. Le TALN connaît des applications multiples : des chatbots de service client aux logiciels de dictée, de la traduction automatisée aux conversations avec Siri, il est partout. Plus particulièrement, la traduction de langues étrangères illustre les progrès phénoménaux enregistrés ces dernières années par cette technologie. Par le passé, les résultats d'une traduction en ligne pouvaient être très aléatoires, mais ils présentent aujourd'hui un degré de fiabilité beaucoup plus élevé. La traduction automatisée n'est toutefois pas encore parfaite et n'est pas encore prête à prendre la place des interprètes. L'un des traducteurs automatiques, Google Translate, a fait l'actualité lorsqu'une capture d'écran traduisant les phrases anglaises « I am smart » et « I am beautiful » vers l'espagnol et le français au masculin (« Je suis intelligent » et « Je suis beau »), et la phrase « I am beautiful and not smart » au féminin (« Je suis belle mais

20 Wijeratne, Y. (2019a), « The Social Media Block isn't helping Sri Lanka », *Slate*. Disponible à l'adresse : https://slate.com/technology/2019/04/sri-lanka-social-media-block-disinformation.html et Wijeratne, Y. (2019b), « Big Tech is as Monolingual as Americans », Foreign Policy. Disponible à l'adresse : https://foreignpolicy.com/2019/05/07/big-tech-is-as-monolingual-as-americans/

21 Manthorpe, R. (2019), « Police share "shooting" video with Facebook to help identify live-streamed attacks », *SkyNews*. Disponible à l'adresse : https://news.sky.com/story/police-share-shooting-video-with-facebook-to-help-identify-live-streamed-attacks-11843511

pas intelligente »), a été partagée[22]. Ces défauts doivent être corrigés et des activités de recherche et développement sont actuellement en cours. Une nouvelle technique, le « masking », créée par l'entreprise chinoise Baidu, permet à un programme de traduction d'aller au-delà de la traduction au mot à mot pour tenir compte du contexte, ce qui produit des résultats plus fiables[23]. Cette technique pourrait s'avérer utile dans le contexte des derniers rapports relatifs au mouvement d'extrême droite Boogaloo, qui utiliserait un langage codé en ligne pour éviter les suppressions automatiques des réseaux sociaux[24].

Ces technologies sont susceptibles de contribuer largement à la modération de contenu sur les sites Internet ayant une conception extrême de la liberté d'expression. Parmi les exemples connus, citons 8chan, 4chan et Gab pour les idéologies d'extrême droite et de nombreuses autres formes de misanthropie liées à des groupes, comme l'antisémitisme, la xénophobie et la suprématie blanche. L'absence de politique peut favoriser l'essor des environnements radicaux, puisqu'elle permet de tout partager à l'exception des contenus illicites tels que la pédopornographie, en vertu de la loi des États-Unis. Les responsables des fusillades de la synagogue de Poway, du Walmart d'El Paso et de la mosquée de Christchurch en 2019 avaient tous posté des publications sur 8chan avant de commettre leurs attaques. Des recherches plus approfondies doivent être menées, mais ces publications finales se démarquent des plaisanteries, des remarques ironiques et du langage très grossier généralement utilisé sur ces sites. Elles mentionnent toutes d'autres attentats et partagent des liens vers un manifeste, une diffusion en direct ou d'autres écrits ; les tireurs affirment qu'ils pourraient perdre la vie. Ils adoptent un ton chaleureux. Il est concevable qu'un système de TALN puisse repérer un niveau de menace croissant en présence de certains indicateurs dans une publication. Ces indicateurs devraient être ajustés aux caractéristiques spécifiques de la plateforme concernée.

Le TALN pourrait aussi créer de la résilience au sein des communautés virtuelles utilisant des langues minoritaires en fournissant de meilleurs services de modération des contenus. La modération automatique de contenu pour les langues des minorités ne parvient pas à produire des résultats fiables. Il serait peut-être plus judicieux, plutôt que d'espérer que les algorithmes produiront bientôt de meilleurs résultats alors que les incitations économiques sont trop faibles pour contraindre les entreprises à investir dans des mises à niveau, de se tourner vers le TALN et l'amélioration des traductions vers les langues parlées par des modérateurs expérimentés et formés. La modération de contenu pourrait ainsi couvrir d'autres langues moins courantes si la traduction automatique est d'un niveau acceptable[25]. Les applications potentielles doivent cependant toujours respecter les normes de protection de la vie privée et les droits humains.

22 Marta Ziosi, LinkedIn. Disponible à l'adresse : https://www.linkedin.com/posts/marta-ziosi-3342007a_googletranslate-googletranslate-women-activity-6603598322009808896-MQJX
23 Baidu Research (2019), « Baidu's Pre-training Model ERNIE Achieves New NLP Benchmark Record ». Disponible à l'adresse : http://research.baidu.com/Blog/index-view?id=128
24 Owen, T. (2020), « The Boogaloo Bois are all over Facebook », *Vice*. Disponible à l'adresse : https://www.vice.com/en_us/article/7kpm4x/the-boogaloo-bois-are-all-over-facebook
25 Wijeratne, Y. (2019b).

3.3 Contenu fictif créé par IA – comment inverser la vapeur

Les contenus manipulés sont susceptibles de permettre à la pensée extrémiste de se répandre dans le discours dominant et peuvent faciliter la radicalisation et entraîner des violences dans le monde réel. La désinformation politique n'est pas une stratégie nouvelle, mais la possibilité d'atteindre des publics d'une ampleur sans précédent d'un simple clic pour aiguiller le débat public engendre de nouvelles difficultés. Ce chapitre cherche des solutions pour lutter contre les trolls, les bots, les fake news et les deepfakes.

Les **trolls** ou les **bots** sont des comptes de médias sociaux qui diffusent des contenus spécifiques ou produisent une mobilisation artificielle sur les plateformes correspondantes. « Ces bots peuvent être programmés pour effectuer des tâches normalement associées aux interactions humaines, comme suivre des utilisateurs, favoriser des tweets, envoyer des messages directs à d'autres utilisateurs et, plus important, tweeter du contenu et retweeter les publications d'un ensemble d'utilisateurs donné ou contenant un hashtag spécifique. »[26] La programmation d'un bot ne nécessite pas de connaissances techniques approfondies et peut être réalisée facilement à l'aide de manuels disponibles en ligne[27].

Les trolls utilisés en grand nombre sont appelés réseau ou armée de trolls ou de bots. Les contenus manipulés orchestrés à l'aide de nombreux messagers peuvent influencer le discours ou les attitudes du public afin de servir un programme particulier. Un exemple bien connu est l'interférence russe dans l'élection présidentielle américaine de 2016, à l'occasion de laquelle de faux engagements placés stratégiquement ont favorisé la campagne de Donald Trump et attaqué Hillary Clinton. Selon les estimations, 5 à 15 % des comptes en ligne étaient faux (ces chiffres sont contestés)[28]. D'après une recherche de Pew Research, les 500 bots les plus actifs sur Twitter sont responsables de 22 % des liens tweetés, contre seulement 6 % environ pour les 500 humains les plus actifs. En parallèle, 66 % des comptes partageant des liens vers les sites Internet les plus populaires sont des bots, contre 34 % de comptes réellement humains[29]. Le besoin d'une réglementation régissant les fermes à trolls a été récemment démontré par l'enquête de Katarzyna Pruszkiewicz[30], qui a travaillé pendant six mois pour une entreprise de trolls polonaise. Elle et ses collègues avaient pour rôle d'orienter, contre rémunération, les conversations en ligne au profit de certains clients, y compris des radio-diffuseurs publics. Il est difficile de savoir dans quelle mesure la mobilisation en ligne se traduit en votes dans la vie réelle[31], mais il est inacceptable, dans une démocratie, que des responsables politiques et des institutions publiques gagnent grâce à l'argent.

26 Symantec Security Response (2018), « How to Spot a Twitter Bot », Symantec Blogs/Election Security. Disponible à l'adresse : https://www.symantec.com/blogs/election-security/spot-twitter-bot
27 Agarwal, A. (2017), « How to write a Twitter Bot in 5 Minutes », Digital Inspiration. Disponible à l'adresse : https://www.labnol.org/internet/write-twitter-bot/27902/
28 Burns, J. (2018), « How many Social Media Users are Real People? », *Gizmodo*. Disponible à l'adresse : https://gizmodo.com/how-many-social-media-users-are-real-people-1826447042
29 Wojcik, S. *et al.* (2017), « Bots in the Twittersphere », Pew Research Centre. Disponible à l'adresse : https://www.pewresearch.org/internet/2018/04/09/bots-in-the-twittersphere/
30 Davies, C. (2019), « Undercover reporter reveals life in a Polish troll farm », *The Guardian*. Disponible à l'adresse : https://www.theguardian.com/world/2019/nov/01/undercover-reporter-reveals-life-in-a-polish-troll-farm
31 Eckert, S. *et al.* (2019), « Die Like Fabrik », *Sueddeutsche Zeitung*. Disponible à l'adresse : https://www.sueddeutsche.de/digital/paidlikes-gekaufte-likes-facebook-instagram-youtube-1.4728833

Selon Mark Zuckerberg, P.-D.G. de Facebook, l'IA est la solution idéale en matière de modération de contenu, puisqu'elle permet l'identification et la suppression des fausses mobilisations de toute nature. D'après lui, seuls les systèmes automatisés peuvent traiter les contenus postés par des millions d'utilisateurs dans différentes langues et issus de différents milieux culturels[32]. Certains détails restent toutefois à préciser. Zuckerberg a également admis, lors de son audition par le Sénat américain en 2018, que l'IA ne serait prête à détecter les nuances linguistiques que d'ici 5 à 10 ans, mais que les évolutions techniques nécessaires à cette fin n'existaient pas encore[33]. Les solutions technologiques existantes prétendent que l'identification des bots est possible. L'hypothèse est la suivante : un bot conçu à des fins spécifiques ne créerait ou ne traiterait que de contenus thématiques uniques ou très restreints, contrairement aux utilisateurs humains, qui s'intéresseraient à un vaste ensemble de sujets. La date et le moment de création du compte figurent également parmi les autres informations à prendre en compte dans l'analyse[34]. La promesse d'une telle technologie s'oppose aux conclusions du Centre d'excellence de l'OTAN de Riga. Ses dernières recherches, qui portent sur Facebook, Twitter, Instagram et YouTube, ont montré que l'identification et la suppression d'engagements fictifs étaient insuffisantes[35]. Pour seulement 300 euros, les chercheurs ont pu acheter 3 530 commentaires, 25 750 likes, 20 000 vues et 5 100 followers. La plateforme n'a pas réussi à classer les comportements ou comptes fictifs : quatre semaines après l'achat, quatre des cinq publications étaient encore en ligne. Même après avoir signalé un échantillon, 95 % des contenus étaient encore en ligne trois semaines après notification des sites Internet concernés. Compte tenu de la détermination des acteurs à diffuser des contenus malveillants à l'aide de faux comptes ou d'armées de trolls, les plateformes doivent adopter une approche proactive pour identifier les faux comptes et éviter ainsi que la modération de contenu se transforme en jeu du chat et de la souris.

Les fake news (fausses informations) contiennent des contenus créés de toutes pièces, des informations carrément fausses ou des théories du complot, qui ne sont pas nécessairement illégaux mais qui posent de vrais problèmes. Une terminologie plus différenciée nous permet de faire une distinction entre désinformation et mésinformation. La première est diffusée intentionnellement, contrairement à la deuxième. Elles sont toutes deux susceptibles de répandre des pensées extrémistes dans le discours dominant, ce qui peut faciliter la radicalisation et entraîner la commission de violences dans le monde réel. La désinformation peut faire partie d'une stratégie politique, et influencer le discours public si elle est diffusée de façon efficace (par exemple par les faux comptes et armées de bots mentionnés plus haut). Par exemple, « Pizzagate » est une théorie du complot issue de la campagne électorale américaine de 2016. Après la fuite des e-mails privés de John Podesta, alors directeur de campagne de la candidate

32 Cao, S. (2019), « Facebook's AI Chief Explains How Algorithms Are Policing Content – And Whether It Works », *The Observer*. Disponible à l'adresse : https://observer.com/2019/12/facebook-artificial-intelligence-chief-explain-content-moderation-policy-limitation/

33 Harwell, D. (2018), « AI will solve Facebook's most vexing problems, Mark Zuckerberg says. Just don't ask when or how », *The Washington Post*. Disponible à l'adresse : https://www.washingtonpost.com/news/the-switch/wp/2018/04/11/ai-will-solve-facebooks-most-vexing-problems-mark-zuckerberg-says-just-dont-ask-when-or-how/

34 Gupta, S. (2017), « A Quick Guide to Identify Twitterbots Using AI », Hackernoon. Disponible à l'adresse : https://hackernoon.com/a-quick-guide-to-identify-twitterbots-using-ai-c3dc3a7b817f

35 Bay, S. et Fredheim R. (2019), « Falling Behind: How social media companies are failing to combat inauthentic behaviour online », NATO STRATCOM COE. Disponible à l'adresse : https://www.stratcomcoe.org/how-social-media-companies-are-failing-combat-inauthentic-behaviour-online

démocrate à la présidentielle, Hillary Clinton, ses opposants ont fait courir le bruit que ses e-mails contenaient des codes rattachant de hauts représentants du parti démocrate à des réseaux de trafic d'êtres humains et de pédophilie. Les partisans d'extrême droite en particulier ont diffusé cette théorie sur les imageboards 4chan et 8chan ainsi que sur Reddit et Twitter pendant la campagne électorale. Plusieurs restaurants ont été cités comme ayant facilité les machinations des prétendus pédophiles. Les propriétaires et employés de ces restaurants ont fait l'objet de menaces, y compris de mort. Finalement, Edgar Maddison Welch, inspiré par les publications en ligne, a décidé de se rendre à l'un des restaurants. Il a tiré trois coups de feu. Personne n'a été blessé. Lors de son interrogatoire après la fusillade, il a nié que les informations étaient fausses.

Bien que le contenu ne soit pas illégal, les fournisseurs de plateformes peuvent pénaliser les violations des normes communautaires qu'ils ont eux-mêmes définies. Pourtant, le filtrage des fausses informations peut aller à l'encontre du modèle commercial des entreprises de médias sociaux. Les contenus polarisants et spectaculaires renforcent la mobilisation des utilisateurs et les poussent à passer plus de temps sur leurs sites, ce qui permet aux entreprises de recueillir toujours plus de données. Ces données représentent le pilier de leur modèle financier, puisqu'elles améliorent la publicité ciblée et accroissent les rendements. Différents acteurs déploient néanmoins des efforts pour lutter contre les fake news. Des chercheurs de l'Université de Waterloo, au Canada, ont développé un outil d'IA qui peut appuyer la vérification des faits à un niveau sans précédent. En recoupant les affirmations d'un article avec d'autres sources, le système indique si elles sont susceptibles ou non d'être des fake news. Selon les chercheurs, ce système a raison neuf fois sur dix[36]. Il pourrait s'agir d'une avancée importante dans la lutte contre les fausses informations.

Newsguard, une initiative du secteur privé dirigée par le géant de la technologie Microsoft, est un bon exemple d'inaptitude. Newsguard est un module d'extension pour navigateurs web qui évalue la crédibilité des médias sous forme de notation. Sur les médias sociaux, il affiche une petite vignette pour indiquer le degré de véracité des informations. Il s'agit d'une solution peu pratique : l'utilisateur doit proactivement télécharger le module d'extension, qui n'aide pas à évaluer des articles précis mais n'attribue qu'une note globale au média à l'origine du support. Ce module n'aurait pas aidé dans l'affaire Pizzagate mentionnée plus haut, diffusée via des comptes privés. Breitbart, un média diffusant une idéologie d'extrême droite et prônant la suprématie blanche, et la chaîne de propagande russe RT reçoivent tous deux une vignette générale verte, suivie toutefois d'une alerte selon laquelle ces sites présentent des « limitations importantes ». Facebook a également été sous le feu des critiques à cause de Breitbart : le réseau social a lancé une section « informations », dans laquelle il publie des histoires tirées de médias vérifiés. Ces médias ont été identifiés en collaboration avec des journalistes et adhèrent aux directives internes de Facebook contre les discours de haine et les contenus « putaclic ». L'inclusion de Breitbart dans les médias en question a provoqué un mouvement de contestation. Facebook a toujours défendu sa décision

36 Grant, M. (2019), « New tool uses AI to flag fake news for media fact-checkers », *Waterloo News*. Disponible à l'adresse : https://uwaterloo.ca/news/news/new-tool-uses-ai-flag-fake-news-media-fact-checkers

en s'appuyant sur l'argument de la liberté d'expression. Pendant ce temps, Twitter a annoncé, au lendemain des élections législatives britanniques de 2019, une interdiction des publicités politiques[37].

En général, l'interdiction des contenus extrémistes joue en faveur de l'hypothèse de l'instauration d'un espace en ligne plus sain grâce à la réduction des possibilités de se retrouver face à des contenus pouvant faciliter la radicalisation. Cette approche ne pourra cependant toujours constituer qu'une partie de la réponse aux contenus préjudiciables en ligne, puisqu'elle ne s'intéresse pas aux causes sous-jacentes de l'opinion exprimée.

Les **deepfakes**, ou vidéos manipulées, sont une version extrême des données de synthèse manipulées et donnent une signification fondamentalement nouvelle à l'expression « faire dire à quelqu'un ce qu'il n'a pas dit ». Les dernières évolutions technologiques permettent aux utilisateurs de créer des vidéos à partir de la voix et des expressions du visage de quelqu'un. Cela peut donner lieu à des vidéos de responsables politiques qui semblent très réalistes sur les plans à la fois de l'image et du son, alors que ces personnes n'ont jamais prononcé ces mots. Un homme politique indien a utilisé une vidéo manipulée pour répondre aux besoins de l'environnement multilingue lors d'une campagne électorale récente, suscitant ainsi des réactions contrastées[38]. Les organisations ayant tiré la sonnette d'alarme à propos des deepfakes ont également diffusé leurs conclusions, comme la vidéo dans laquelle Boris Johnson et son opposant Jeremy Corbyn s'accordent mutuellement un soutien politique pour l'élection de décembre 2019. Ces données créées de toutes pièces sont particulièrement présentent dans le secteur de la pornographie, faisant des femmes les plus grandes victimes de ces nouvelles technologies : les femmes peuvent désormais être les stars de vidéos pornographiques à leur insu et sans y avoir consenti. Cette technologie semble particulièrement néfaste lorsqu'elle est combinée à des outils de **désidentification**. Ces outils modifient les images et vidéos de telle manière qu'il est impossible pour les algorithmes d'identifier une nouvelle version légèrement modifiée comme le visage d'origine. Ils les identifient plutôt comme un nouvel élément à part entière. Les utilisateurs peuvent toutefois reconnaître le visage d'origine dans la version modifiée. Ces technologies peuvent faire obstacle à une suppression rapide et efficace. Le Forum mondial de l'Internet contre le terrorisme (GIFCT), dirigé par le secteur, a créé un Hash-Sharing Consortium (consortium de partage de hash) – une base de données d'« empreintes digitales » numériques des contenus malveillants, ou « hash »[39]. Il souhaite renforcer son efficacité en assurant la collaboration entre différentes sociétés[40]. On ne sait pas très bien si la base de données pourrait résister à l'utilisation

37 La différenciation entre publicité politique et publicité thématique demeure contestée et il n'existe aucune définition universellement reconnue. Les difficultés survenues ont donné naissance à des appels pour traiter toutes les publicités selon une norme stricte afin de renforcer la transparence et de permettre un contrôle sur leur impact. Pour plus d'informations, voir : Frederik J. Zuiderveen Borgesius *et al.* (2018), « Online Political Microtargeting: Promises and Threats for Democracy ». Utrecht Law Review, 14 (1). 82-96. Disponible à l'adresse : https://www.ivir.nl/publicaties/download/UtrechtLawReview.pdf ; et « Universal Advertising Transparency by default » (2020). Disponible à l'adresse : https://epd.eu/wp-content/uploads/2020/09/joint-call-for-universal-ads-transparency.pdf

38 Christopher, N. (2020), « We've just seen the First Use of Deepfakes in an Indian Election Campaign », *Vice*. Disponible à l'adresse : https://www.vice.com/en_in/article/jgedjb/the-first-use-of-deepfakes-in-indian-election-by-bjp

39 GIFCT, Joint Tech Innovation. Disponible à l'adresse : https://www.gifct.org/joint-tech-innovation/

40 Llansó, E. (2019), « Platforms want centralised Censorship. That should scare you », *Wired*. Disponible à l'adresse : https://www.wired.com/story/platforms-centralized-censorship/ ; et Windwehr, S. et York, J. (2020), « One Database to rule them all: The invisible Content Cartel that undermines the freedom of expression online », EFF. Disponible à l'adresse : https://www.eff.org/deeplinks/2020/08/one-database-rule-them-all-invisible-content-cartel-undermines-freedom-1.

systématique de logiciels de désidentification, notamment en cas de diffusion stratégique de contenus à caractère extrémiste via un ensemble d'acteurs et de plateformes.

Pour être réalistes, les deepfakes nécessitent des connaissances approfondies – en particulier si le résultat vise à tromper le public. Les connaissances techniques nécessaires empêchent encore une intensification rapide de la technologie, notamment lorsqu'il existe d'autres méthodes moins complexes pouvant atteindre le même objectif. Par ailleurs, comme l'indique Hwang, les outils de désidentification s'accompagnent d'un risque d'exposition médiatique. Les politiques d'interdiction et le risque d'exposition médiatique peut rendre les deepfakes moins attractifs pour influencer les campagnes[41]. Twitter a utilisé pour la première fois ses nouvelles vignettes relatives aux contenus manipulés sur des contenus créés par le directeur des réseaux sociaux de la Maison-Blanche[42]. Sa politique prévoit un signalement par la plateforme de toutes les vidéos ou photos manipulées sans qu'elles soient supprimées pour autant, à moins que le contenu représente une menace pour la sécurité d'une personne.

La maîtrise de la désinformation et de la mobilisation artificielle nécessite un renforcement de la culture numérique des utilisateurs. Comme le montrent les recherches sur la diffusion de fausses informations sur Twitter, les mensonges se propagent plus rapidement et plus largement que la vérité, et cela est dû aux interactions humaines. Les bots renforcent l'aspect viral mais ne sont pas à l'origine de la vaste propagation des mensonges. Les chercheurs mettent une telle diffusion sur le compte des réactions émotionnelles et de la relative nouveauté des contenus[43]. Leurs conclusions montrent clairement qu'il n'y a pas d'autre solution qu'une éducation adaptée permettant aux utilisateurs de faire preuve de plus de résilience lorsqu'ils surfent sur le net.

41 Hwang, T. (2020), « Deepfakes – A grounded threat assessment », Centre for Security and Emerging Technology.
42 Dent, S. (2020), « Twitter labels video retweeted by Trump as "manipulated data" », Engadget Online. Disponible à l'adresse : https://www.engadget.com/2020/03/09/twitter-labels-trump-retweet-manipulated-media/
43 Vosoughi, S. *et al.* (2018), « The spread of true and false news online », *Science* vol. 359, n° 6380, p. 1146–51. Disponible à l'adresse : https://science.sciencemag.org/content/359/6380/1146

4 Prédire la radicalisation – une IA générale au service des forces de l'ordre

Il est facile d'imaginer la pièce où les décisions sont prises ; empruntons l'image du classique de science-fiction *Minority Report* : un grand écran bleu tactile affichant les résultats d'une machine superintelligente censée assister les forces de l'ordre. Le résultat affiché est basé sur les données disponibles des individus ainsi que sur leur comportement en ligne en temps réel. Les alarmes du système retentissent dès que ce dernier repère une hausse considérable des facteurs de risque témoignant d'un niveau inquiétant de radicalisation. En fonction du comportement observé, le système envoie alors les unités compétentes cueillir le suspect. Ce système d'IA fiable permet à la police d'agir avant même qu'une infraction ne soit commise grâce à son pouvoir prédictif. Ce scénario peut sembler tentant, même s'il est exagéré et qu'il relève davantage de la science-fiction que de la réalité. Le lien entre nouvelles technologies et sécurité fait toutefois naître des aspirations pour ce type d'avancées. Ce chapitre porte uniquement sur le mythe d'une IA générale douée d'une intelligence supérieur, qui aurait pour but de surveiller le contenu publié par les individus et leur comportement en ligne pour lutter contre la radicalisation.

Les projets de police prédictive s'intéressent à la façon dont l'IA peut prêter assistance aux forces de l'ordre dans leur travail. Ces projets sont des applications d'AA qui prédisent la commission future d'infractions à partir de corrélations statistiques pour venir en aide aux forces de l'ordre[44]. L'efficacité de ces systèmes soulève de vives controverses. Par exemple, la police du comté de Kent, au Royaume-Uni, a cessé d'utiliser le logiciel américain PredPol pour prédire la commission d'infractions, compte tenu des doutes concernant sa valeur ajoutée[45]. L'organisation de défense des libertés civiles Big Brother Watch a signalé un usage excessif de la force à l'encontre des minorités et la réactivation de préjugés à l'encontre de certains quartiers dans les projets de police prédictive, puisque la présence accrue de patrouilles dans les quartiers historiquement plus exposés à la criminalité entraîne également un signalement accru des infractions et crée un cycle de renforcement des biais structurels[46]. Le recours aux indicateurs pour prédire la criminalité nécessite un raisonnement interprétatif et des formes non causales de réflexion sur les risques. Cette évolution reflète une tendance à mettre l'accent sur l'importance du contexte dans l'analyse des risques et peut être considérée comme un pas de plus vers le profilage, perçu comme

44 Moses, B. L. et Chan, J. (2018), « Algorithmic prediction in policing: assumptions, evaluation, and accountability », *Policing and Society*, vol. 28, n° 7, p. 806–22.
45 Big Brother Watch (2019), « Big Brother Watch Submission to the Centre for Data Ethics and Innovations: Bias in Algorithmic Decision Making (Crime and Justice) ». Disponible à l'adresse : https://bigbrotherwatch.org.uk/wp-content/uploads/2019/06/Big-Brother-Watch-submission-to-the-Centre-for-Data-Ethics-and-Innovation-Bias-in-Algorithmic-Decision-Making-Crime-and-Justice-June-2019.pdf
46 Ibid.

injuste et discriminatoire dans de nombreuses sociétés[47]. Le profilage racial ou ethnique a mis en lumière les rapports entre forces de l'ordre et sociétés pluralistes[48].

Cette approche fondée sur des indicateurs pourrait toutefois paraître tentante dans le cadre des efforts déployés en matière de lutte contre la radicalisation en ligne. Une liste d'indicateurs basés sur le comportement en ligne et les contenus consommés pourrait servir de fondement pour appuyer la recherche d'individus en cours de radicalisation. Il est toutefois difficile d'imaginer le transfert des systèmes actuels de police prédictive vers les efforts en matière de lutte contre la radicalisation, et ce pour trois raisons.

Le premier obstacle est le manque de clarté ou de compréhension précise des processus de radicalisation[49]. Il est par ailleurs impossible de définir avec précision le moment du passage à l'acte d'un individu radicalisé, qui justifierait une intervention. La radicalisation et le terrorisme ne sont heureusement pas suffisamment répandus pour fournir un ensemble de données fiable. La radicalisation est un processus très complexe et individualisé, et bien que les chercheurs aient identifié certains éléments récurrents dans les différents processus de radicalisation[50], les informations sont insuffisantes pour entraîner un algorithme. Les systèmes actuels d'IA ont besoin d'une grande quantité de données pour développer leur pouvoir de prédiction. En l'absence de percée technologique permettant aux technologies d'IA de travailler avec un nombre beaucoup moins important de données, la situation ne semble guère prometteuse. Il n'existe à l'heure actuelle aucune indication d'une telle avancée.

Les systèmes de police prédictive actuels fonctionnent sur la base d'hypothèses portant sur des groupes, combinées à des informations sur des lieux et moments relatifs à des zones exposées à la criminalité. En d'autres termes, pour faire des prédictions éclairées, les algorithmes se fondent sur un ensemble d'informations en libre accès, de données gouvernementales et de données fournies par des entreprises privées. Les hypothèses sous-jacentes sont fondées sur des choix beaucoup plus économiques et rationnels. Prenons l'exemple d'un cambriolage : si un cambrioleur a réussi son coup à une certaine heure et dans un certain quartier, il est susceptible de recommencer à la même heure et dans un quartier similaire pour mettre toutes les chances de son côté. L'idée est la suivante : les malfaiteurs veulent minimiser le risque et maximiser leurs chances de succès autant que possible. Ces hypothèses ne fonctionnent pas forcément pour la radicalisation et le terrorisme. Cela ne veut pas dire qu'il n'y a pas de raisonnement rationnel derrière la voie choisie du terrorisme, mais ce raisonnement ne fonctionne pas de la même manière que les délits d'appropriation. Mourir pour la bonne cause a au contraire été un facteur d'attraction dans la campagne de propagande de l'EI pour rejoindre le califat : « On ne meurt qu'une fois – pourquoi ne pas mourir en martyr ? »[51].

47 Monaghan, J. et Molnar, A. (2016), « Radicalisation theories, policing practices, and "the future of terrorism?" », *Critical Studies on Terrorism*, vol. 9, n° 3, p. 393–413.
48 Open Society Foundations (2019), « Ethnic Profiling: What it is and Why it must end ». Disponible à l'adresse : https://www.opensocietyfoundations.org/explainers/ethnic-profiling-what-it-and-why-it-must-end
49 Voir Monaghan et Molnar.
50 Voir Neumann.
51 Kingsley, P. (2014), « Who is behind Isis's terrifying online propaganda operation? », *The Guardian*. Disponible à l'adresse : https://www.theguardian.com/world/2014/jun/23/who-behind-isis-propaganda-operation-iraq

La troisième raison est la limitation dans les démocraties libérales découlant de l'idée qu'un individu est protégé par l'État et de l'État. Réfléchissons : de quoi aurions-nous besoin pour prédire le comportement d'un individu ? Un algorithme prédisant un comportement individuel aurait besoin de données plus différenciées que l'information disponible fondée sur le groupe. En d'autres termes, pour garantir la fiabilité de ses prédictions, il aurait besoin de données non anonymisées sur le comportement des individus – et plus il y en a, mieux c'est. Cela supposerait une surveillance sans précédent des comportements individuels : il faudrait alors mettre en place un suivi en temps réel de la société tout entière. Ce n'est ni compatible avec les droits à la protection de la vie privée en vigueur, ni souhaitable dans une société libre, pour des raisons éthiques et morales. Cela aurait des répercussions sur les droits fondamentaux, comme les libertés d'expression, de la presse ou d'association, le caractère privé des télécommunications, etc.[52] En bref, cela ne ferait rien d'autre que donner corps à une véritable dystopie.

52 Ganor, B. (2019), « Artifical or Human: A New Era of Counterterrorism Intelligence? », *Studies in Conflict and Terrorism*.

Pour aller plus loin

Une IA démocratique par nature

Les algorithmes fonctionnent à l'aide de données, et leur soif de données est insatiable. Ils doivent dans un premier temps être entraînés à l'aide d'un vaste ensemble de données, puis testés à l'aide de données supplémentaires pour continuer à analyser sans relâche toujours plus de données. Il n'est pas surprenant que les questions de respect de la vie privée et de protection des données viennent à l'esprit, notamment en conjonction avec l'IA et les problèmes de sécurité.

Au lieu d'essayer de réguler les systèmes automatisés de prise de décision pour répondre aux normes des sociétés démocratiques, les valeurs démocratiques devraient être incorporées aux technologies dès leur conception. Les développements technologiques doivent être « privés par défaut », c'est-à-dire traiter les données des utilisateurs en respectant les normes les plus rigoureuses en matière de protection de la vie privée, à moins que les utilisateurs acceptent de partager leurs informations. Ceci a des répercussions pour la conservation des contenus sur les médias sociaux et le suivi de masse des données comportementales personnelles. De plus, les systèmes doivent produire des résultats transparents ou fournir des explications, ce qui permet aux opérateurs humains d'évaluer les calculs de l'algorithme et de décider de la crédibilité des résultats. Cela pourrait s'avérer être une alternative claire à l'actuel « Blackbox AI », dont les résultats ne peuvent être expliqués. Une meilleure transparence par exemple dans le domaine des systèmes de recommandation ou de conservation des contenus faciliterait par ailleurs l'intérêt public et la recherche, ce qui entraînerait à son tour une meilleure compréhension du phénomène de radicalisation en ligne. La responsabilité dans les processus décisionnels ne peut être atteinte que grâce à une IA transparente et digne de confiance. Les audits d'applications décisionnelles automatisées assurerait une mise en œuvre licite et assurerait la mise en place d'incitatifs pour une IA juste et démocratique.

5 Conclusion

Le présent rapport visait à discuter de la façon dont les technologies basées sur l'IA peuvent assister la lutte contre la radicalisation en ligne.

L'IA offre de nouvelles opportunités d'analyser les mégadonnées et de prédire l'avenir. Il serait bon d'assurer une application étroite de la technologie pour appuyer la prévention de l'extrémisme violent et réduire le risque pour les internautes de tomber sur des contenus susceptibles d'entraîner une radicalisation. Les moteurs de recherche, les systèmes de recommandation et le TALN sont des outils fondés sur l'IA particulièrement prometteurs. Le TALN présente un potentiel pour une meilleure modération des contenus en ligne, notamment en ce qui concerne les langues parlées uniquement par de petits groupes de population. Souvent, les rendements financiers supposés incitant les grandes plateformes à investir dans la modération de contenu des langues minoritaires – notamment des modérateurs humains – ne sont pas suffisamment attractifs. Les plateformes plus modestes n'ont pas toujours l'expertise technique ou les ressources nécessaires pour se doter de systèmes de modération des contenus, puisque même l'utilisation des modèles existants suppose beaucoup de temps et d'efforts. D'autres plateformes défendant une interprétation extrême de la liberté d'expression affirment ne pas vouloir restreindre leurs utilisateurs. Un TALN amélioré peut aider à traduire des contenus dans des langues parlées par un plus grand nombre de modérateurs formés. Il peut également détecter les modèles sémantiques inhabituels sur les sites ne souhaitant pas investir dans la modération de contenu. Toutefois, ces mesures doivent toujours garantir le respect des normes de confidentialité et des droits humains.

La modération de contenu pour les grandes plateformes de médias sociaux demeure problématique. Le grand nombre de langues, associé à des contextes culturels très variés, est encore trop insurmontable pour permettre aux algorithmes de filtrer les contenus malveillants. Un vaste débat public sur les contenus entrant dans la « zone grise » – c'est-à-dire les contenus malveillants mais licites – s'impose. La société doit trouver un terrain d'entente sur les limites de la liberté d'expression en ligne. Cette décision ne devrait pas revenir qu'aux entreprises privées. De même, les technologies d'IA sont sous-développées et inutiles pour lutter contre la mobilisation artificielle en ligne, tel que les trolls, les bots et les fake news, dont la détection laisse encore à désirer à l'heure actuelle. Les internautes doivent être invités à participer et éduqués pour garantir des comportements responsables en ligne conduisant à une souveraineté numérique ; la conception des plateformes devrait permettre la mise en place d'un système transparent de « notification et action ». Les consommateurs et utilisateurs ne doivent pas être les seuls à porter le poids de cette responsabilité. La recherche d'une sécurité en ligne doit être appuyée par des stratégies qui dissuadent la publication de contenus malveillants ou fictifs en ligne et interdire les modèles commerciaux priorisant les contenus préjudiciables, puisque cela augmente la mobilisation en ligne et profite aux recettes publicitaires. L'essentiel sera pour les opérateurs de plateformes d'éviter de supprimer trop de

contenu et d'empiéter sur la liberté d'expression tout en fournissant des mesures adaptées pour prévenir la publication de contenus malveillants.

Il est peut-être devenu évident qu'une IA générale, c'est-à-dire un système doué d'une intelligente supérieure, n'est pas une solution pour prédire la radicalisation en ligne des individus, et ce pour deux raisons. Tout d'abord, une raison technique : compte tenu de l'état actuel des technologies d'IA, les algorithmes ont besoin de vastes quantités de données pour faire des prédictions utiles. Heureusement, la radicalisation et le terrorisme ne sont pas suffisamment répandus pour produire suffisamment de données pour qu'une IA générale puisse prédire le comportement d'individus en matière de radicalisation en ligne. Le taux de faux positifs et de faux négatifs serait insupportable. Des investissements dans les ressources humaines seraient plus bénéfiques. La deuxième raison concerne le respect de la vie privée : un système observant les comportements en ligne en temps réel, stockant et analysant les données, ne respecterait pas les normes de protection de la vie privée en vigueur dans les démocraties libérales. La mise en place d'un tel système pourrait entraîner une surveillance de masse de la société.

À long terme, l'application de technologies fondées sur l'IA devra suivre des normes précises. Ces normes viseront à protéger les utilisateurs contre les outils de prise de décision automatique injustes, par exemple parce qu'ils se fondent sur des ensembles de données biaisés ou sur une conservation des données discriminatoire basée sur le genre, le sexe, la religion ou d'autres caractéristiques protégées par la loi sur les droits des personnes. Les résultats des algorithmes doivent être atteints de façon transparente pour assurer une responsabilité des décisions fondées sur des calculs algorithmiques. Le développement d'une IA qui respecte le droit à la vie privée et qui soit non discriminatoire et compréhensible pour l'opérateur doit être la voie à suivre.

Contexte politique

Cette section a été rédigée par Armida van Rij et Lucy Thomas, toutes deux adjointes de recherche au Policy Institute du King's College, à Londres. Elle fournit un aperçu du contexte politique dans lequel s'inscrit ce rapport.

Introduction

Empêcher la glorification de la violence et du terrorisme, la diffusion de la désinformation et d'autres formes de contenus extrémistes en ligne constitue un défi que doivent relever les acteurs politiques et plateformes technologiques du monde entier. Le rapport introductif *Intelligence artificielle (IA) et lutte contre l'extrémisme violent (LEV)*, destiné au Forum mondial de l'Internet contre le terrorisme (GIFCT), donne un aperçu global des opportunités et difficultés posées par l'IA dans le contexte de la LEV.

La pression pesant sur les décideurs politiques nationaux et internationaux, ainsi que sur les sociétés technologiques, pour modérer et supprimer plus rapidement et efficacement les contenus extrémistes est de plus en plus forte. Cela est dû en partie au fait que les préjudices commis dans la vie réelle, tragiques de par leur fréquence et leur ampleur, comme la fusillade de mars 2019 dans la mosquée de Christchurch, celle commise dans l'église de Charleston, aux États-Unis, en 2015, ou encore celle touchant la mosquée de Québec, au Canada, en 2017, ont été commis à la suite de la publication de contenus malveillants en ligne. Le secteur des technologies et les décideurs politiques nationaux et multinationaux ont depuis pris des mesures pour supprimer un ensemble de contenus extrémistes produits par l'État islamique et d'autres groupes djihadistes violents, ainsi que des contenus misogynes, antisémites, islamophobes ou prônant la suprématie blanche.

Le rapport introductif *Intelligence artificielle et lutte contre l'extrémisme violent* décrit des technologies d'IA conçues pour aider, accélérer ou rendre plus précises la modération et la suppression de contenu en ligne. Cela va des outils « entraînés » sur le plan linguistique pour identifier et signaler les contenus nuisibles aux technologies qui détectent les vidéos manipulées (deepfakes), en passant par les processus d'apprentissage automatique visant à élaborer des outils d'identification algorithmique. Le rapport cite également plusieurs difficultés et obstacles au déploiement effectif de ces technologies. Tout d'abord, les systèmes de recommandation fondés sur l'IA peuvent emporter les internautes dans des spirales constituées de contenus de plus en plus néfastes. Deuxièmement, l'accent placé sur la modération des contenus dans les langues européennes entraîne une reconnaissance et une modération sous-optimales des contenus diffusés dans des langues non dominantes. Troisièmement, il n'existe à l'heure actuelle aucune mesure efficace pour lutter contre la désinformation en ligne, que ce soit sur le plan technique ou éthique. Enfin, un système général d'IA pour suivre les échanges en ligne en temps réel se fonderait sur un nombre démesuré de données non anonymisées, ce qui mettrait en péril les droits à la protection de la vie privée et à la liberté d'expression.

Dans ce rapport, nous observons les mesures prises par neuf acteurs politiques nationaux et supranationaux clés pour aborder ces opportunités et difficultés : le Canada, la France, le Japon, le Ghana, la Nouvelle-Zélande, le Royaume-Uni, les États-Unis, la Commission européenne et la Direction exécutive du Comité contre le terrorisme des Nations Unies. Nous présentons un aperçu de ces efforts au cas par cas et concluons en présentant des recommandations stratégiques.

Intelligence artificielle et lutte contre l'extrémisme violent : relever les défis et évaluer les nouvelles avancées

Canada

Le gouvernement canadien a cultivé une stratégie rigoureuse en matière de lutte contre le terrorisme, et ses efforts et initiatives en matière de LEV en ligne ne constituent qu'une partie d'une politique holistique plus vaste de LEV. Ses investissements dans la lutte contre l'extrémisme violent en ligne, et l'attention qu'il y porte, sont malheureusement la conséquence d'actes préjudiciables commis dans la vie réelle, comme c'est le cas pour de nombreux autres gouvernements.

Fin janvier 2017, un Québécois du nom d'Alexandre Bissonnette a ouvert le feu sur le Centre culturel islamique de la ville de Québec, faisant six morts et cinq blessés. L'enquête ouverte par la suite a découvert que Bissonnette avait joué un rôle actif dans des cercles racistes et d'extrême droite en ligne avant la fusillade, et consultait régulièrement les comptes Twitter de théoriciens du complot, de nationalistes blancs et de personnalités de la droite alternative comme Ben Shapiro et Alex Jones d'InfoWars[53].

Contrairement aux auteurs de nombreux autres attentats terroristes très médiatisés et facilités par les activités sur la toile, Bissonnette n'a pas publié de manifeste ou de déclaration d'intention en ligne[54]. Le recours croissant aux manifestes terroristes est néanmoins une tendance globale que l'IA peut aider à combattre. Les manifestes d'extrême droite se renvoient souvent les uns aux autres, leurs auteurs exprimant par exemple leur admiration pour des attentats perpétrés plus ou moins récemment ou reproduisant des mèmes ou raccourcis Internet. L'IA pourrait aider à repérer le téléchargement de contenus préjudiciables, tels que des manifestes d'extrême droite, en vue d'assurer une intervention préalablement à la commission d'attentats dans la vie réelle.

53 Riga, A. (17 avril 2018), « Quebec Mosque Killer Confided He Wished He Had Shot More People, Court Told », *Montreal Gazette*. Disponible à l'adresse : https://montrealgazette.com/news/local-news/quebec-mosque-shooter-alexandre-bissonnette-trawled-trumps-twitter-feed/. Voir également : Mahrouse, G. (2018), « Minimizing and denying racial violence: Insights from the Quebec Mosque shooting », *Revue Femmes et Droit*, vol. 30, n° 3, p. 471–93.

54 Par exemple, Robert Bowers (auteur de la fusillade de 2018 dans une synagogue de Pittsburgh), Dylann Roof (2015, fusillade dans une église de Charleston) ; Brenton Tarrant (2019, fusillade dans la mosquée de Christchurch), Patrick Crusius (2019, fusillade d'El Paso), Anders Breivik (2011, massacre d'Utøya), et de nombreux autres ont posté des manifestes sur diverses plateformes en ligne peu de temps avant de passer à l'acte. Voir : Ware, J. (2020), *Testament to Murder: The Violent Far-Right's Increasing Use of Terrorist Manifestos* – Note d'information, Centre international de lutte contre le terrorisme – La Haye. Disponible à l'adresse : https://icct.nl/publication/testament-to-murder-the-violent-far-rights-increasing-use-of-terrorist-manifestos/

La réponse du Canada à l'extrémisme violent en ligne, telle qu'énoncée dans sa Stratégie nationale de lutte contre la radicalisation menant à la violence[55], a trois objectifs : concevoir des contre-discours en collaboration avec la société civile, appuyer la recherche relative à la LEV et développer des partenariats avec des initiatives internationales et des sociétés technologiques. Le troisième objectif, en particulier, est l'espace dans lequel le Canada a investi concernant le lien entre IA et LEV.

Plus particulièrement, en 2019, le Canada a confié à Tech Against Terrorism, une initiative internationale parrainée par les Nations Unies travaillant avec le secteur mondial des technologies, la mission de développer la Plateforme d'analyse des contenus à caractère terroriste (TCAP)[56], une base de données qui héberge des supports et contenus terroristes vérifiés provenant de sources libres de droits et d'ensembles de données existants. Le but recherché est que la TCAP agira comme service d'alerte en temps réel en cas de publication de contenus à caractère terroriste et extrémiste violent sur les plateformes Internet plus modestes : le contenu malveillant vérifié publié sur ces plateformes sera rapidement partagé avec les équipes de modération de contenu, qui prendront promptement des mesures à leur encontre. À moyen et long terme, la TCAP servira d'archive historique pour une analyse académique quantitative et qualitative améliorée[57].

Dans le domaine de l'IA plus précisément, l'un des objectifs déclarés de la TCAP est d'appuyer le développement d'un écosystème de classificateurs algorithmiques de contenus[58]. Comme le montre le rapport introductif du GNET, *Intelligence artificielle et lutte contre l'extrémisme violent*, « les algorithmes ont besoin de vastes quantités de données pour faire des prédictions utiles »[59]. Les mécanismes de modération automatisée de contenu fondés sur l'apprentissage automatique et le traitement automatique du langage naturel s'appuient sur l'analyse de mégadonnées pour entraîner l'IA à reconnaître les données pour ce qu'elles sont et les éléments de vidéos de propagande de l'EI (logos, drapeaux, etc.) pour identifier et marquer les vidéos futures présentant des éléments similaires ou identiques. La TCAP, en sa qualité de première plateforme unifiée consacrée au contenu à caractère terroriste en ligne, est une véritable mine d'or de données et informations pour les développeurs concevant des algorithmes d'apprentissage automatique pour repérer et classer les supports à caractère terroriste.

En fournissant des contenus à caractère terroriste vérifiés provenant de différentes plateformes en ligne sous forme d'archive historique, la TCAP pourrait réaliser une avancée technologique significative dans la lutte contre l'extrémisme violent en ligne. Le gouvernement canadien, en sa qualité de coparrain de la plateforme, a montré combien les investissements ciblés et intelligents dans des initiatives transsectorielles pouvaient donner à la recherche, à l'industrie et à la société civile la possibilité de collaborer pour faire avancer l'IA dans le contexte de la LEV.

---

55 « Stratégie nationale de lutte contre la radicalisation menant à la violence », Sécurité publique Canada. Disponible à l'adresse : https://www.securitepublique.gc.ca/cnt/rsrcs/pblctns/ntnl-strtg-cntrng-rdclztn-vlnc/index-fr.aspx

56 La TCAP est également mentionnée dans la partie « Contexte politique » du rapport du GNET intitulé « Décrypter la haine : emploi de l'analyse de texte expérimentale aux fins de classification des contenus à caractère terroriste ». Disponible à l'adresse : https://gnet-research.org/wp-content/uploads/2020/09/GNET-Report-Decoding-Hate-Using-Experimental-Text-Analysis-to-Classify-Terrorist-Content_FRENCH.pdf

57 « Press release: Tech Against Terrorism Participates in UN General Assembly Week in New York », Tech Against Terrorism. Disponible à l'adresse : https://www.techagainstterrorism.org/2019/10/08/press-release-tech-against-terrorism-participates-in-un-general-assembly-week-in-new-york/

58 Ibid.

59 Schroeter, M. (2020), « Intelligence artificielle et lutte contre l'extrémisme violent : rapport introductif », Global Network on Extremism and Technology, p. 28.

Commission européenne

Dans son Livre blanc sur l'intelligence artificielle de février 2020, la Commission européenne indique que « l'IA peut fournir des outils permettant de mieux protéger les Européens contre les actes criminels et terroristes »[60]. L'approche de l'UE concernant l'utilisation de l'IA se veut à la fois normative et axée sur les investissements. Elle vise notamment à favoriser l'émergence d'une « IA digne de confiance » en élaborant un cadre réglementaire solide pour protéger tous les Européens et contribuer à la « création d'un marché intérieur fluide » en vue du développement de l'IA[61]. « Pour pouvoir être dignes de confiance, les systèmes d'IA (…) doivent être robustes et précis sur le plan technique »[62]. L'UE prévoit également de multiplier les investissements dans l'IA pour atteindre au moins 20 milliards d'euros par an d'ici 2030[63].

La Commission européenne a nommé un Groupe d'experts de haut niveau sur l'intelligence artificielle (AI HLEG) en 2019. Ce groupe a défini sept séries de critères pour garantir une IA digne de confiance. Ces sept principes sont les suivants : facteur humain et contrôle humain ; robustesse et sécurité ; respect de la vie privée et gouvernance des données ; transparence ; diversité, non-discrimination et équité ; bien-être sociétal et environnemental ; responsabilisation[64]. Dans cette logique, la Commission appelle, dans son Livre blanc, à la mise en place d'un « écosystème de confiance » permettant le respect des droits fondamentaux[65].

Le Livre blanc sur l'IA a reçu une réaction mitigée de la part des grandes sociétés technologiques. Google a appelé l'UE à exploiter les règlements et cadres réglementaires existants plutôt que d'en élaborer de nouveaux auxquels les sociétés technologiques devraient se plier. En parallèle, Google, Facebook et d'autres plateformes technologiques devront se préparer à se conformer à la loi relative aux services numériques, prévue cette année, qui visera à « réguler l'écosystème en ligne dans toute une série de domaines, y compris … les contenus choquants »[66]. Dans le sillage de son Livre blanc sur l'IA, l'UE devrait adopter, plus tard cette année, de nouvelles législations sur l'IA et la sécurité, la responsabilité, les droits fondamentaux et les données[67].

60 Commission européenne, (19 février 2020), « Livre blanc sur l'intelligence artificielle – Une approche européenne axée sur l'excellence et la confiance », p. 2. Disponible à l'adresse : https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_fr.pdf
61 Ibid., p. 12.
62 Ibid., p. 24.
63 Gouvernement français, ministère de l'Europe et des Affaires étrangères, « Transparence et responsabilité, les enjeux de l'intelligence artificielle ». Disponible à l'adresse : https://www.diplomatie.gouv.fr/fr/politique-etrangere-de-la-france/diplomatie-numerique/transparence-et-responsabilite-les-enjeux-de-l-intelligence-artificielle/
64 Ibid.
65 Ibid.
66 Stolton, S. (23 juin 2020), « Platform clamp down on hate speech in run up to Digital Services Act », *EURACTIV*. Disponible à l'adresse : https://www.euractiv.com/section/digital/news/platforms-clamp-down-on-hate-speech-in-run-up-to-digital-services-act/
67 Kayali, L., Heikkila, M. et Delcker, J. (19 février 2020), « Europe's digital vision, explained », *Politico*. Disponible à l'adresse : https://www.politico.eu/article/europes-digital-vision-explained/

France

En France, l'IA relève des attributions de plusieurs acteurs clés. Le coordonnateur ministériel en matière d'intelligence artificielle a pour mission d'analyser et de développer des propositions sur les mutations liées à la transformation numérique dans le domaine de la sécurité intérieure[68]. L'Agence de l'innovation de défense du ministère de la Défense compte également une cellule de coordination de l'intelligence artificielle de défense.

La France a adapté son cadre juridique de façon à assurer une utilisation sûre et plus efficace des technologies d'IA afin de protéger sa population. Sur le plan des politiques, la France a publié sa stratégie en matière d'IA en mars 2018. Ses principaux objectifs sont les suivants : améliorer le système d'éducation et de formation à l'IA pour développer et attirer les meilleurs talents ; adopter une politique d'open data pour la mise en œuvre des applications d'IA et la mutualisation des actifs ; et développer un cadre éthique pour une utilisation transparente et juste des applications d'IA[69]. Conformément à la directive de l'UE sur la sécurité des réseaux et des systèmes d'information, la France a élaboré sa propre loi en matière de cybersécurité[70]. Elle est par ailleurs actuellement en train de synthétiser ses réflexions sur l'usage des technologies d'IA à des fins militaires[71].

La France a lancé une série d'initiatives axées sur l'IA. Lors de la Conférence multipartite du G7 sur l'intelligence artificielle de 2018, la France et le Canada ont annoncé le lancement d'un Groupe international d'experts en IA ayant pour mission d'appuyer l'adoption responsable de systèmes d'IA[72]. Les deux pays, rejoints ensuite par plusieurs autres, ont également lancé le développement d'un nouveau Partenariat mondial pour l'intelligence artificielle (PMIA). L'objet de cette initiative est d'orienter le développement et l'utilisation responsables de l'IA, en tenant compte des questions relatives aux droits de la personne, à l'inclusion, à la diversité, à l'innovation et à la croissance économique[73]. Elle servira plus précisément à combler l'écart entre théorie et pratique sur l'IA en appuyant la recherche consacrée aux activités liées à l'IA. En France, le PMIA est soutenu par un centre d'excellence, une institution sœur-du Centre d'excellence du PMIA à Montréal. Il bénéficie également du soutien de l'OCDE.

Conformément à sa stratégie relative à l'IA, la France pourra envisager d'adopter une loi pour une République numérique, qui servirait à « ouvrir les données publiques, renforcer la protection des droits des utilisateurs et de la confidentialité des données, et garantir que les opportunités dues à la numérisation bénéficient à tous »[74].

---

68 Gouvernement français, Bureau du Premier ministre, (13 juillet 2018), « Plan d'action contre le terrorisme », p. 20. Disponible à l'adresse : http://www.sgdsn.gouv.fr/uploads/2018/07/plan-d-action-contre-le-terrorisme-v8.pdf

69 Commission européenne, *France AI strategy report*. Disponible à l'adresse : https://ec.europa.eu/knowledge4policy/ai-watch/france-ai-strategy-report_en

70 Gouvernement français, Agence nationale de la sécurité des systèmes d'information, *Directive network and information system security (NIS)*. Disponible à l'adresse : https://www.ssi.gouv.fr/entreprise/reglementation/directive-nis/

71 Pannier, A. et Schmitt, O. (2019), « To fight another day: France between the fight against terrorism and future warfare », *International Affairs* vol. 95, n° 4. Disponible à l'adresse : https://academic.oup.com/ia/article/95/4/897/5492774

72 Gouvernement du Canada, Bureau de Premier ministre (6 décembre 2018), *Mandat pour le Groupe international d'experts en intelligence artificielle*. Disponible à l'adresse : https://pm.gc.ca/fr/nouvelles/notes-dinformation/2018/12/06/mandat-groupe-international-dexperts-intelligence

73 Gouvernement français, ministère de l'Europe et des Affaires étrangères (15 juin 2020), *Lancement du Partenariat Mondial sur l'Intelligence Artificielle par quinze membres fondateurs*. Disponible à l'adresse : https://www.diplomatie.gouv.fr/fr/politique-etrangere-de-la-france/diplomatie-numerique/actualites-et-evenements/article/lancement-du-partenariat-mondial-sur-l-intelligence-artificielle-par-quinze

74 Commission européenne, *France AI strategy report*.

Ghana

Les efforts pour lutter contre l'extrémisme violent en ligne sont limités au Ghana, puisque la violence politique dans le pays n'est pas due à des activités terroristes, contrairement à ses voisins le Nigéria et le Tchad[75]. La Global Terrorism Database, une base de données sur les attentats terroristes perpétrés dans le monde depuis 1970, ne recense que 21 incidents et 23 victimes en 50 ans au Ghana[76].

Le Ghana n'est pas confronté aux mêmes problématiques que certains de ses voisins, concernant notamment les blocages d'Internet par les pouvoirs publics ou l'exploitation des réseaux sociaux par le gouvernement pour éliminer l'opposition politique[77]. Ces gouvernements ont exploité l'héritage laissé par les lois coloniales, historiquement utilisées pour violer les libertés et « justifier de nombreuses (…) tentatives de requêtes extrajudiciaires adressées au secteur privé »[78]. Le rapport « Ranking Digital Rights » de 2019 indique que les médias sociaux et les fournisseurs d'accès à Internet ont dû répondre à des requêtes extrajudiciaires de blocage de la part des gouvernements, suscitant des inquiétudes concernant une surveillance et une censure excessives[79].

Bien que le gouvernement ghanéen n'ait pas encore fait de telles requêtes illégales, les groupes de la société civile et les journalistes se sont dits inquiets pour l'avenir[80]. Avant les élections de 2016, le chef de la police ghanéenne a annoncé un éventuel blocage des réseaux sociaux[81]. Bien que le président ait résisté à ces projets, les préoccupations relatives aux droits numériques au Ghana grandissent.

Les lois libérales relatives à la liberté d'expression au Ghana laissent la place aux abus sur les espaces numériques, comme les discours haineux et le harcèlement en ligne (en particulier à l'encontre des femmes)[82]. Les appels à une régulation plus stricte des médias sociaux vont donc croissant. Selon un expert de la Fondation pour les Médias en Afrique de l'Ouest, « en l'absence de réglementation, d'autres dispositions législatives », comparables aux requêtes gouvernementales décrites plus haut, « seront exploitées pour poursuivre les gens en justice de manière potentiellement excessive ».

La régulation des médias sociaux par le gouvernement doit, toutefois, rechercher un équilibre entre protection des utilisateurs et protection de leur liberté d'expression. Un groupe de la société civile influent qui fait campagne contre les blocages d'Internet a lancé l'alerte contre les réglementations gouvernementales en matière de médias sociaux :

75 Merci à Tomiwa Ilori, chercheur à l'Unité sur la liberté d'expression et d'information et les droits numériques du Centre des droits de la personne de l'Université de Prétoria, pour ces informations obtenues par courriel.
76 Global Terrorism Database, START. Disponible à l'adresse : https://www.start.umd.edu/gtd/
77 Ilori, T. (2020), « Moderate Globally Impact Locally: Content Moderation Is Particularly Hard in African Countries », Information Society Project at Yale Law School. Disponible à l'adresse : https://law.yale.edu/isp/initiatives/wikimedia-initiative-intermediaries-and-information/wiii-blog/moderate-globally-impact-locally-content-moderation-particularly-hard-african-countries
78 Ilori, T. (2020), « Stemming digital colonialism through reform of cybercrime laws in Africa », Information Society Project at Yale Law School. Disponible à l'adresse : https://law.yale.edu/isp/initiatives/wikimedia-initiative-intermediaries-and-information/wiii-blog/stemming-digital-colonialism-through-reform-cybercrime-laws-africa
79 Ranking Digital Rights, « 2019 RDR Corporate Accountability Index ». Disponible à l'adresse : https://rankingdigitalrights.org/index2019/assets/static/download/RDRindex2019report.pdf
80 Majama, K. (2019), « Africa in urgent need of a homegrown online rights strategy », Association for Progressive Communications. Disponible à l'adresse : https://www.apc.org/en/news/africa-urgent-need-homegrown-online-rights-strategy
81 Olukotun, D. (16 août 2019), « President of Ghana says no to internet shutdowns during coming elections », *AccessNow*. Disponible à l'adresse : https://www.accessnow.org/president-ghana-says-no-internet-shutdown-elections-social-media/
82 Endert, J. (2018), « Digital backlash threatens media freedom in Ghana », DW Akademie. Disponible à l'adresse : https://www.dw.com/en/digital-backlash-threatens-media-freedom-in-ghana/a-46602904

« Si nous permettons au gouvernement de réglementer Internet – et il existe des exemples dans d'autres pays – celui-ci finira par nous dire comment l'utiliser »[83].

On ignore encore si le Ghana a pour projet de développer des outils d'IA pour faciliter la régulation des contenus en ligne. Les menaces régionales pesant sur la liberté d'expression, héritées des lois de l'époque coloniale, ont toutefois montré que le pays doit placer les droits numériques de ses citoyens au cœur de tout outil technologique ou effort législatif pour contrôler les contenus préjudiciables en ligne. Dans le cadre d'une initiative louable, le Ghana a adopté, en 2019, un projet de loi sur le droit à l'information, qui garantit l'accès à l'information détenue par les institutions publiques[84]. Ce projet de loi signale que le gouvernement ghanéen souhaite faire preuve de transparence et de responsabilité dans sa gestion des droits numériques. Toute évolution future concernant la modération des contenus en ligne doit suivre ces normes et engagements afin de protéger les droits à la vie privée et à la liberté d'expression.

Japon

Le Japon déploie la plupart de ses efforts en matière de LEV par l'intermédiaire de l'Association des nations de l'Asie du Sud-Est (ANASE)[85]. Dès 2004, les États membres de l'ANASE, en partenariat avec le Japon, ont publié un ensemble de déclarations de coopération en matière de lutte contre le terrorisme international. Ces déclarations, outre le fait qu'elles signalaient des intentions politiques, engageaient les signataires à « prévenir, empêcher et combattre le terrorisme international grâce à l'échange d'informations, le partage de renseignements et le renforcement des capacités », établissant ainsi un précédent en matière de coopération régionale pour la lutte contre l'extrémisme violent et le terrorisme[86].

En 2015, le Japon a réaffirmé son engagement en faveur d'une collaboration internationale en Asie du Sud-Est pour la lutte contre l'extrémisme violent et le terrorisme et d'une coopération dans la mise en œuvre du Plan d'action 2018-2025 de l'ANASE pour prévenir et contrer l'essor de la radicalisation et de l'extrémisme violent[87]. Ce plan d'action accorde la priorité aux partenariats avec « le monde des affaires et le secteur des technologies pour promouvoir la modération et améliorer le dialogue en vue de prévenir la radicalisation et l'extrémisme violent », ainsi qu'au renforcement des « communications stratégiques » pour empêcher l'utilisation malveillante des médias sociaux par les groupes extrémistes et terroristes[88].

---

83 Ibid.

84 Yahya Jafru, M. (26 mars 2019), « Right to information – RTI bill passed into law », *Graphic Online*. Disponible à l'adresse : https://www.graphic.com.gh/news/politics/ghana-news-rti-bill-passed.html

85 « Japan: Extremism & Counter Extremism », Counter-Extremism Project. Disponible à l'adresse : https://www.counterextremism.com/countries/japan

86 « ASEAN-Japan Joint Declaration for Cooperation to Combat International Terrorism », ANASE. Disponible à l'adresse : https://asean.org/?static_post=asean-japan-joint-declaration-for-cooperation-to-combat-international-terrorism-2

87 ANASE (2015), « Chairman's Statement of the 18th ASEAN-Japan Summit, Kuala Lumpur, November 22, 2015 ». Disponible à l'adresse : https://www.asean.org/wp-content/uploads/2015/12/6.-Chairmans-Statement-of-the-18th-ASEAN-Japan-Summit-Final-Final.pdf ; Mission du Japon à l'ANASE, « Japan's cooperation with ASEAN 2025 (Counter-Terrorism) ». Disponible à l'adresse : https://www.asean.emb-japan.go.jp/asean2025/jpasean-ps03.html

88 ANASE (2018), « 2018 ASEAN Plan of Action to Prevent and Counter the Rise of Radicalisation and Violent Extremism (2018–2025), adopted in Myanmar, October 31, 2018 ». Disponible à l'adresse : https://cil.nus.edu.sg/wp-content/uploads/2019/10/2018%20ASEAN%20Plan%20of%20Action%20to%20Prevent%20and%20Counter%20the%20Rise%20of%20Radicalisation%20and%20Violent%20Extremism%20(2018-2025).pdf

Les Jeux Olympiques et Paralympiques de 2020 (repoussés à 2021 à cause de la crise du coronavirus) auront lieu à Tokyo. L'organisation des J.O. a toujours été considérée comme une « mise à l'essai » des capacités d'un pays en matière de sécurité, et les jeux seront l'occasion pour le Japon de tester ses innovations en matière d'IA, de sécurité et d'exécution des lois[89].

À l'occasion d'un test mené en amont des Jeux en 2018, la police de la préfecture de Kanagawa a annoncé la mise en place d'un système de police prédictive fondé sur un algorithme d'apprentissage automatique profond visant à prévenir les infractions et attentats[90]. Depuis, les plus grandes sociétés technologiques ont confirmé la fourniture de systèmes de reconnaissance faciale, d'authentification biométrique et de détection des comportements à grande échelle aux Jeux, ainsi qu'aux ports et aéroports nippons[91]. Ces systèmes auront la capacité de scanner les visages à la recherche d'émotions particulières et de confirmer l'identité des individus à partir de leurs traits du visage et d'informations personnelles.

On ignore encore si ces capacités en matière de sécurité fondée sur l'IA s'étendra aux médias sociaux et aux activités en ligne. Les systèmes testés à Kanagawa auraient pu inclure la surveillance des contenus publiés sur les médias sociaux pour lutter contre la criminalité, ce qui pourrait être perçu comme une validation par les forces de l'ordre japonaises du suivi par IA des médias sociaux pour lutter contre les contenus malveillants et potentiellement dangereux en ligne. Une telle démarche serait risquée, compte tenu des protestations qui ont fusé en 2017 contre le projet de loi controversé du gouvernement relatif à la lutte contre le terrorisme, considéré par ses détracteurs comme mettant en péril les libertés civiles[92]. Le Japon devra accorder une attention particulière à la protection des droits à la vie privée et à la liberté d'expression de ses citoyens dans le cadre du développement de technologies d'IA pour lutter contre l'extrémisme violent en ligne et hors ligne.

Nouvelle-Zélande

En Nouvelle-Zélande, la lutte contre l'extrémisme violent en ligne est régie par de nombreux organismes et agences qui travaillent en collaboration. Citons par exemple le Comité du Cabinet chargé des relations extérieures et de la sécurité ; les agences de communication des services de police, de renseignement et de sécurité ; et les agences chargées des affaires étrangères, du commerce, de la défense, du transport, de l'innovation et du développement. La stratégie globale de la Nouvelle-Zélande en matière de lutte contre le terrorisme est présentée dans son plan de stratégie nationale publié en février 2020[93].

89 Voir, par exemple : Soria, V. (2011), « Beyond London 2012: The Quest for a Security Legacy », *The RUSI Journal*, vol. 156, n° 2, p. 36-43.

90 « Kanagawa police to launch AI-based predictive policy system before Olympics », *Japan Times* (29 janvier 2018). Disponible à l'adresse (accès payant) : https://www.japantimes.co.jp/news/2018/01/29/national/crime-legal/kanagawa-police-launch-ai-based-predictive-policing-system-olympics/

91 Gouvernement du Japon (2019), « All is Ready for a Safe and Secure Tokyo Games ». Disponible à l'adresse : https://www.japan.go.jp/tomodachi/2019/autumn-winter2019/tokyo2020.html ; « NEC Becomes a Gold Partner for the Tokyo 2020 Olympic and Paralympic Games », NEC Corporation (2015). Disponible à l'adresse : https://www.nec.com/en/press/201502/global_20150219_01.html

92 « Japan passes controversial anti-terror conspiracy law », *BBC* (15 juin 2017). Disponible à l'adresse : https://www.bbc.co.uk/news/world-asia-40283730

93 Gouvernement de la Nouvelle-Zélande, Comité des fonctionnaires chargés de coordonner la sécurité intérieure et extérieure, Comité de coordination de la lutte contre le terrorisme (février 2020), « Countering terrorism and violent extremism national strategy overview ». Disponible à l'adresse : https://dpmc.govt.nz/sites/default/files/2020-02/2019-20 CT Strategy-all-final.pdf

Après la fusillade de la mosquée de Christchurch en mars 2019, les gouvernements de la Nouvelle-Zélande et de la France ont formé une coalition de chefs d'État et de sociétés technologiques et de médias sociaux dans le cadre de l'Appel de Christchurch pour supprimer les contenus terroristes et extrémistes violents en ligne[94]. Les États signataires s'engagent dans le cadre de cet appel à assurer l'application des lois interdisant la diffusion de contenus à caractère terroriste et extrémiste violent en ligne tout en respect le droit international des droits de l'homme, y compris la liberté d'expression. Les pays œuvrent également pour appuyer les cadres et activités de renforcement des capacités et de sensibilisation visant à empêcher l'utilisation des services numériques à des fins de diffusion de contenus à caractère terroriste et extrémiste violent.

L'Appel de Christchurch engage également les sociétés, y compris Amazon, Facebook, Google, Twitter et YouTube, à adopter des normes sectorielles plus strictes en matière de responsabilité et de transparence. Lesdites sociétés doivent assurer l'exécution de leurs normes communautaires et conditions de service en accordant la priorité aux mesures de modération et de suppression de contenu et en identifiant les contenus en temps réel pour examen et évaluation. Ensemble, les pays et sociétés déploient des efforts en conjonction avec la société civile pour promouvoir les activités communautaires afin d'intervenir dans les processus de radicalisation en ligne.

Une Commission d'enquête royale a été mise en place au lendemain de l'attentat de mars 2019 pour évaluer la réaction des agences face à la fusillade et déterminer quelles mesures supplémentaires pouvaient être prises pour empêcher de futures attaques[95]. Le rapport de la Commission mettra en lumière la stratégie actuelle de la Nouvelle-Zélande et ses orientations futures en matière de LEV, et permettra de comprendre le rôle joué par l'IA à l'avenir. La publication du rapport a été repoussée à l'hiver 2020 en raison de la crise du coronavirus.

Le gouvernement de la Nouvelle-Zélande s'astreint par ailleurs à des normes de transparence et de responsabilité plus rigoureuse dans son utilisation des algorithmes à des fins de gouvernance. Comme le fait remarquer l'auteure de *Intelligence artificielle et lutte contre l'extrémisme violent : rapport introductif*, l'utilisation des algorithmes peut exacerber les préjugés[96]. En juillet 2020, le gouvernement a publié la Charte des algorithmes pour la Nouvelle-Zélande (Algorithm Charter for Aotearoa New Zealand), un examen exhaustif de l'utilisation des algorithmes par l'État dans les secteurs allant des transports à la justice, ainsi qu'un engagement pour une meilleure transparence, une mobilisation plus forte des parties prenantes, une meilleure protection du droit à la vie privée et une supervision humaine plus approfondie de l'utilisation des algorithmes[97]. Cette charte – la première en son genre – a été signée, au moment de la rédaction de ce rapport, par 25 organismes publics.

94 Voir https://www.appeldechristchurch.com/

95 The Royal Commission of Inquiry into the Attack on Christchurch mosques. Voir : https://christchurchattack.royalcommission.nz/

96 Voir également Babuta, A. et Oswald, M. (2019), « Briefing Paper: Data Analytics and Algorithmic Bias in Policing », RUSI. Disponible à l'adresse : https://www.gov.uk/government/publications/report-commissioned-by-cdei-calls-for-measures-to-address-bias-in-police-use-of-data-analytics ; Benjamin, R. (2019), *Race After Technology: Abolitionist Tools for the New Jim Code* (Polity) ; Benjamin, R., « A New Jim Code? », Berkman Klein Center for Internet & Society at Harvard University. Enregistrement disponible à l'adresse : https://cyber.harvard.edu/events/new-jim-code

97 « Algorithm charter for Aotearoa New Zealand », data.govt.nz. Disponible à l'adresse : https://data.govt.nz/use-data/data-ethics/government-algorithm-transparency-and-accountability/algorithm-charter

Les agences et organismes responsables de la LEV en ligne manquent toutefois à l'appel. Le rôle réservé par les décideurs politiques de Nouvelle-Zélande à l'IA et aux outils algorithmiques pour lutter contre la publication de contenus malveillants en ligne et les normes régissant ces outils demeurent donc encore incertains. La charte représente un pas dans la bonne direction, et l'application de ces normes aux mesures de LEV fondées sur l'IA constituerait une évolution positive en matière d'élaboration de politiques.

Royaume-Uni

En février 2018, le Royaume-Uni a annoncé le développement d'un outil algorithmique basé sur l'apprentissage automatique visant à détecter les contenus à caractère terroriste publiés en ligne par l'État islamique. Ledit logiciel était entraîné à repérer et signaler les éléments audiovisuels reconnaissables de la propagande de l'EI – drapeaux, logos, format, structures et bandes sonores. Au fil des ans, les grandes plateformes technologiques comme YouTube et Facebook ont beaucoup investi dans le développement de leurs propres outils de modération automatisée de contenu. L'outil britannique était conçu pour être utilisable sur n'importe quelle plateforme et était donc à la libre disposition des plateformes Internet et médias sociaux plus modestes, tels que Vimeo.

Bien que prometteur, il présente une efficacité restreinte et a reçu un accueil très limité. Tout d'abord, comme l'a souligné notre collègue Charlie Winter, l'EI publie des contenus très variables en ligne, tels que vidéos, photos, écrits ou encore bulletins radiophoniques. Bien que la lutte contre les contenus vidéos soit une démarche positive, « cela ne permettra, au mieux, que d'atténuer quelque peu [le problème], sans constituer toutefois une véritable solution »[98]. Deuxièmement, le ministère de l'Intérieur a commandé cet outil pour reconnaître les contenus vidéos les plus explicites et choquants de l'EI. La société de développement de logiciels qui a conçu l'outil a expliqué qu'il s'agissait « moins d'une question de volume, [et qu'il était] davantage question du rôle qu'il [le ministère de l'Intérieur] pensait que le logiciel pouvait jouer pour lutter contre un ensemble particulier de vidéos »[99]. Toutefois, de nombreuses études académiques ont décrit la grande incidence que pouvait avoir la propagande « plus douce » et le potentiel de radicalisation qu'elle pouvait avoir sur des périodes plus longues[100]. Le fait de limiter de trois façons – EI, vidéos et contenus extrêmes – l'IA à des fonctions étroites fait obstacle à l'efficacité technique de l'outil. Celui-ci a été mis gratuitement à la disposition des sociétés technologiques plus modestes, mais aucune entreprise ne l'avait encore adopté en avril 2019[101].

L'approche adoptée par le gouvernement britannique concernant l'utilisation de l'IA dans la lutte contre l'extrémisme violent en ligne met également en lumière les conflits d'intérêts potentiels pouvant apparaître dans le cadre de la collaboration entre les pouvoirs publics et le secteur. Le Livre blanc du gouvernement sur les dangers en ligne (Online Harms

98 Temperton, J. (13 février 2018), « ISIS could easily dodge the UK's AI-powered propaganda blockade », *Wired*. Disponible à l'adresse : https://www.wired.co.uk/article/isis-propaganda-home-office-algorithm-asi
99 Ibid.
100 « Hashtag Terror: How ISIS Manipulates Social Media », Anti-Defamation League (21 août 2014). Disponible à l'adresse : https://www.adl.org/education/resources/reports/isis-islamic-state-social-media
101 Murgia, M. et Bond, D. (6 avril 2019), « Businesses show no appetite for anti-terror AI tool », *Financial Times*. Disponible à l'adresse (accès payant) : https://www.ft.com/content/fda2d218-56fb-11e9-91f9-b6515a54c5b1

White Paper), publié en avril 2019, a présenté un argumentaire complet pour une meilleure régulation nationale des médias sociaux[102]. En vertu de ce nouveau cadre réglementaire, les sociétés technologiques et médias sociaux auront un nouveau devoir légal de diligence à l'égard de leurs utilisateurs, supervisé par Ofcom, l'organisme britannique de régulation des communications. Ofcom soumettra les plateformes à des sanctions financières et techniques – les sites pourraient être bloqués par les fournisseurs d'accès à Internet et devoir payer une amende s'élevant à 4 % maximum de leur chiffre d'affaires mondial – pour non-respect du cadre et violation du devoir légal de diligence[103]. Tout en annonçant la sortie de l'outil algorithmique en février 2018, la ministre de l'Intérieur de l'époque, Amber Rudd, a indiqué que les entreprises pourraient être légalement obligées de l'adopter.

En soi, ces mesures réglementaires ne soulèvent pas d'inquiétudes. La société de développement de logiciels et d'analyse de données à l'origine de l'outil, ASI Data Science (aujourd'hui Faculty), était toutefois chargée de la modélisation des données dans les campagnes en faveur du retrait de l'UE Vote Leave et Leave.EU, et a donc été impliquée dans le scandale Cambridge Analytica[104]. Par ailleurs, en mai 2020, elle a obtenu au moins sept marchés publics en l'espace de dix-huit mois, et entretient des liens personnels et commerciaux notables avec Dominic Cummings, conseiller spécial du Premier ministre[105].

Ces faits soulèvent des inquiétudes concernant l'existence d'un conflit d'intérêts. La confiance du public et des entreprises dans l'outil est sapée par le fait que le contrat de développement de l'outil a été confié à une entreprise entretenant des liens étroits avec le cercle restreint du gouvernement et impliquée dans un scandale public, et par la défense d'une législation qui obligerait les médias sociaux à l'utiliser pour garantir la continuité de leurs activités.

Une société de développement indépendante du gouvernement aurait pu développer un outil plus efficace sur le plan technique et adapté à des requêtes plus larges (capable de reconnaître d'autres contenus que les seules vidéos publiées par l'EI, par exemple), et renforcer ainsi la confiance et l'adoption de l'outil. La transparence et la responsabilité ne sont « pas de simples termes vides de sens qu'il suffit d'invoquer pour la forme : elles sont essentielles à la réussite des efforts visant à résoudre les problèmes » en matière de lutte contre l'extrémisme violent en ligne à l'aide de l'IA[106]. Le Royaume-Uni a manqué une occasion stratégique importante de développer et fournir des technologies d'IA de pointe pour modérer les contenus préjudiciables en ligne en sapant la confiance en l'outil et en compromettant son efficacité technique.

---

102 Gouvernement britannique (avril 2019), « Online Harms White Paper » Disponible à l'adresse : https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/793360/Online_Harms_White_Paper.pdf

103 Crawford, A. (29 juin 2020), « Online Harms bill: Warning over 'unacceptable' delay », *BBC*. Disponible à l'adresse : https://www.bbc.co.uk/news/technology-53222665

104 Cadwalladr, C. (7 mai 2017), « The great British Brexit robbery: how our democracy was hijacked », *The Guardian*. Disponible à l'adresse : https://www.theguardian.com/technology/2017/may/07/the-great-british-brexit-robbery-hijacked-democracy

105 Evans, R. et Pegg, D. (4 mai 2020), « Vote Leave AI firm wins seven government contracts in 18 months », *The Guardian*. Disponible à l'adresse : https://www.theguardian.com/world/2020/may/04/vote-leave-ai-firm-wins-seven-government-contracts-in-18-months ; Pegg, D., Evans, R. et Lewis, P. (12 juillet 2020), « Revealed: Dominic Cummings firm paid Vote Leave's AI firm £260,000 », *The Guardian*. Disponible à l'adresse : https://www.theguardian.com/politics/2020/jul/12/revealed-dominic-cummings-firm-paid-vote-leaves-ai-firm-260000 ; Pegg, D. et Evans, R. (2 juin 2020), « AI firm that worked with Vote Leave given new coronavirus contract », *The Guardian*. Disponible à l'adresse : https://www.theguardian.com/technology/2020/jun/02/ai-firm-that-worked-with-vote-leave-wins-new-coronavirus-contract

106 « Tackling the Information Crisis: A Policy Framework for Media System Resilience », The Report of the LSE Commission on Truth Trust & Democracy, p. 32. Disponible à l'adresse : https://www.lse.ac.uk/media-and-communications/assets/documents/research/T3-Report-Tackling-the-Information-Crisis-v6.pdf

Direction exécutive du Comité contre le terrorisme des Nations Unies

La Direction exécutive du Comité contre le terrorisme des Nations Unies (UN CTED) a été créée par la Résolution 1535 (2004) du Conseil de sécurité des Nations Unies comme organe spécialisé ayant pour mandat de soutenir le Comité contre le terrorisme du Conseil de sécurité[107]. Son but initial était d'évaluer la mise en œuvre par les États membres des Nations Unies des résolutions du Conseil de sécurité en matière de lutte contre le terrorisme et d'appuyer leurs efforts par le dialogue. L'UN CTED travaille en étroite collaboration avec le Conseil de sécurité, les principales sociétés technologiques et les organisations de la société civile via le GIFCT.

Plusieurs Résolutions du Conseil des Nations Unies portent sur l'utilisation abusive d'Internet à des fins terroristes, et l'UN CTED cherche à améliorer la cohérence et à rationaliser les points d'intersection entre les résolutions du Conseil de sécurité des Nations Unies et le rôle des technologies de l'information. La Résolution 2129 (2013) du Conseil de sécurité reconnaît les liens de plus en plus étroits entre terrorisme et informatique, de même que l'exploitation des technologies comme Internet pour commettre et faciliter des actes terroristes en favorisant l'incitation, le recrutement, la levée de fonds ou la planification d'actes terroristes[108]. Cette résolution renforce également le mandat de l'UN CTED. Les Résolutions 2354 (2017), 2395 (2017) et 2396 (2017) prient les États membres de l'ONU de coopérer pour empêcher les organisations terroristes d'exploiter les TIC, et de travailler avec le secteur privé et la société civile pour élaborer des mesures effectives de prévention de l'utilisation abusive d'Internet à des fins terroristes[109]. La Résolution 1373 du Conseil de sécurité leur demande d'intensifier et d'accélérer « l'échange d'informations opérationnelles » concernant l'usage des technologies informatiques par les organisations terroristes et de mettre un terme au recrutement de membres de réseaux terroristes[110].

Le Groupe de haut niveau sur la coopération numérique du Secrétaire général des Nations Unies recherche des solutions pour atténuer les risques de l'IA[111]. Il propose, dans sa recommandation 3C :

> *« Nous estimons que les systèmes intelligents autonomes devraient être conçus de manière à pouvoir expliquer leurs décisions et à garantir une responsabilisation humaine derrière leur utilisation. Les audits et les systèmes de certification devraient permettre de vérifier la conformité des systèmes d'intelligence artificielle aux normes d'ingénierie et de déontologie, qui devraient s'élaborer en suivant une approche multilatérale et multipartite. Les décisions de vie ou de mort ne devraient pas être déléguées à des machines. Nous appelons au renforcement de la coopération*

107 Chowdhury Fink, N. (2012), « Meeting the Challenge: A guide to United Nations counterterrorism activities », *Institut international de la paix*, p. 45. https://www.ipinst.org/wp-content/uploads/publications/ebook_guide_to_un_counterterrorism.pdf

108 Nations Unies, Comité contre le terrorisme du Conseil de sécurité (14 septembre 2018), « Public-private efforts to address terrorist content online: A year of progress – what's next? ». Disponible à l'adresse : https://www.un.org/sc/ctc/news/event/public-private-efforts-address-terrorist-content-online-year-progress-whats-next/ ; Global Initiative Against Transnational Organised Crime, *Responding to terrorist use of the Internet* (21 janvier 2019). Disponible à l'adresse : https://globalinitiative.net/terrorist_use_internet/

109 Nations Unies, Comité contre le terrorisme du Conseil de sécurité, 2018.

110 Global Initiative Against Transnational Organised Crime, 2019.

111 Gouvernement français, ministère de l'Europe et des Affaires étrangères, « Transparence et responsabilité, les enjeux de l'intelligence artificielle ».

numérique multipartite afin de réfléchir à la conception et à l'application de normes et principes tels que la transparence et la non-partialité dans les systèmes intelligents autonomes dans différents contextes sociaux. »[112]

L'un des domaines prioritaires concerne la protection des droits humains dans l'ère numérique[113].

États-Unis

La LEV en ligne constitue un axe prioritaire de la Stratégie de lutte contre le terrorisme des États-Unis, qui engage les pouvoirs publics à collaborer avec les entreprises et le secteur pour lutter contre les processus de recrutement, de levée de fonds et de radicalisation terroristes en ligne. En ce qui concerne les initiatives transnationales, les États-Unis travaillent avec des initiatives comme Tech Against Terrorism et le Forum mondial de lutte contre le terrorisme, qui s'appuie sur un partenariat avec d'autres signataires, la société civile et le secteur de la technologie pour concevoir des approches de lutte contre l'extrémisme violent en ligne à moyen et long terme.

En ce qui concerne le domaine législatif national, des appels à la régulation des médias sociaux et plateformes technologiques ont été motivés par des cas signalés d'interférence et d'opérations d'information russes en ligne lors de l'élection présidentielle de 2016. En même temps, les entreprises de médias sociaux ont continué à croître, tant sur le plan du nombre d'utilisateurs que sur celui des services et produits subsidiaires proposés. Fin 2019, le Comité sénatorial des banques et la Commission de l'énergie et du commerce de la Chambre des Représentants américaine ont organisé des auditions sur Libra, le projet de cryptomonnaie de Facebook. Ces auditions ont donné l'occasion au législateur américain d'interroger les dirigeants de Facebook sur la manipulation et l'utilisation abusive de leur plateforme[114], et de faire de la régulation des grandes sociétés technologiques une option législative viable[115].

Alors que le Congrès envisage de prendre des mesures législatives, le renseignement américain a quant à lui pris les choses en main en utilisant l'IA pour lutter contre l'extrémisme violent en ligne. Au printemps et à l'été 2019, les États-Unis ont été secoués par une vague de fusillades, dont les auteurs entretenaient depuis longtemps des liens avec l'extrémisme violent en ligne. Par exemple, l'auteur de la fusillade de la synagogue de Poway, en Californie, fin avril 2019, a publié un manifeste sur 8chan peu de temps avant de perpétrer son attentat. Ledit manifeste renvoie à d'autres fusillades provoquées par les activités en ligne, comme la fusillade de la mosquée de

112 ONU (16 décembre 2019), « L'ère de l'interdépendance numérique – Rapport du Groupe de haut niveau sur la coopération numérique créé par le Secrétaire général de l'Organisation des Nations Unies ». Disponible à l'adresse : https://www.un.org/sites/www.un.org/files/uploads/files/Ere_Interdependance_numerique.pdf

113 ONU, Groupe de haut niveau du Secrétaire général sur la coopération numérique. Disponible à l'adresse : https://www.un.org/en/digital-cooperation-panel/

114 Commission de l'énergie et du commerce de la Chambre des Représentants des États-Unis, « Facebook: Transparency and Use of Consumer Data », transcription du 11 avril 2018, p. 33. Disponible à l'adresse : https://docs.house.gov/meetings/IF/IF00/20180411/108090/HHRG-115-IF00-Transcript-20180411.pdf

115 Molla, R. et Stewart, E. (2019), « How 2020 Democrats think about breaking up Big Tech », *Vox*. Disponible à l'adresse : https://www.vox.com/policy-and-politics/2019/12/3/20965447/tech-2020-candidate-policies-break-up-big-tech

Christchurch et celle de la synagogue de Pittsburgh, ainsi qu'à des personnalités et références de l'extrême droite et du mouvement nationaliste blanc en ligne.

Dans ce contexte, le Federal Bureau of Investigation (FBI) a publié un appel d'offres demandant à des sous-traitants privés de développer une technologie donnant au Bureau un « accès en temps quasi-réel à un vaste ensemble d'échanges sur les médias sociaux » pour « repérer l'éventail croissant de menaces pesant sur les intérêts nationaux américains, y mettre un terme et enquêter dessus »[116]. Un appel d'offres similaire a été émis en janvier 2020[117]. En juin 2020, alors que le pays était secoué par le mouvement de protestations #BlackLivesMatter, le FBI a prolongé ses contrats avec Dataminr, une société de suivi et d'analyse des médias sociaux, et Venntel, une entreprise de données de localisation[118].

Ces technologies et cet accès aux données équivaudraient à un système général d'IA, comme l'indique la partie 4 du rapport introductif *Intelligence artificielle et lutte contre l'extrémisme violent*, c'est-à-dire un système prédictif permettant aux forces de l'ordre d'intervenir sur la base d'un mécanisme d'alerte. Ce type d'outils présenterait une menace remarquable sur le plan de la déontologie pour les droits à la vie privée des utilisateurs, puisque la surveillance en temps réel du comportement individuel à des fins d'application de la loi reposerait sur des données non anonymisées. La collecte de données identifiables porterait atteinte aux droits à la sécurité personnelle, à la protection de l'identité et à la liberté d'expression.

L'approche des États-Unis en matière d'utilisation de l'intelligence artificielle pour lutter contre l'extrémisme violent en ligne prouve les difficultés légales et éthiques inhérentes au suivi et à la modération des contenus en ligne. Comme l'écrit Marie Schroeter, une telle approche « ne ferait rien d'autre que donner corps à une véritable dystopie »[119].

116 US Government Federal Acquisitions Service, « Contract Opportunity: Social Media Alerting Subscription ». Disponible à l'adresse : https://beta.sam.gov/opp/b6de554012cf4ab9ab795f52c638467c/view

117 US Government Federal Acquisitions Service, « Request for Proposal – FBI Social Media Alerting ». Disponible à l'adresse : https://beta.sam.gov/opp/2b3003e9b0b34b639687786e8420013b/view

118 US Government Federal Acquisitions Service, « Contract Information – Dataminr. Inc. ». Disponible à l'adresse : https://beta.sam.gov/awards/90552288%2BAWARD?keywords=15F06720P0000950&sort=-relevance&index=&is_active=true&page=1 ; Fang, L. (24 juin 2020), « FBI Expands Ability to Collect Cellphone Location Data, Monitor Social Media, Recent Contracts Show », *The Intercept*. Disponible à l'adresse : https://theintercept.com/2020/06/24/fbi-surveillance-social-media-cellphone-dataminr-venntel/

119 Schroeter, M. (2020), « Intelligence artificielle et lutte contre l'extrémisme violent : rapport introductif », Global Network on Extremism and Technology, p. 25.

Recommandations stratégiques

Les initiatives et actions existantes, telles que celles décrites ci-dessus, fournissent des renseignements, ainsi que des recommandations, aux décideurs politiques du monde entier. Nos conclusions nous amènent à formuler les recommandations stratégiques suivantes :

Recommandation 1 : Créer un organisme de réglementation indépendant à l'échelle transnationale pour superviser les efforts nationaux de lutte contre l'extrémisme violent en ligne fondée sur l'intelligence artificielle

Les normes gouvernementales créant des sanctions pour les médias sociaux manquant à leur devoir de modérer les contenus préjudiciables[120] peuvent être très efficaces[121], mais risquent de restreindre les droits des citoyens à la liberté d'expression, la crainte d'encourir des sanctions pouvant pousser les entreprises à faire de l'excès de zèle. Comme nous l'avons vu précédemment dans les cas du Royaume-Uni, du Japon et des États-Unis, les législations et les mesures relatives à la modération de contenu sont susceptibles de se heurter à des problèmes d'ordre légal et éthique en matière de respect de la vie privée, de confiance et de responsabilité.

L'auto-régulation des médias sociaux, dans le cadre de laquelle les sociétés créent et appliquent leurs propres normes, codes et politiques en matière de suppression de contenu malveillant en ligne, peut s'avérer efficace, mais l'applicabilité des normes peut être irrégulière et opaque[122]. De nombreuses grandes entreprises publient des données de haut niveau sur la modération de contenu, mais elles n'y sont pas obligées[123].

Une corégulation par le gouvernement, la société civile et le secteur des technologies, supervisée par un organisme transnational indépendant, devrait être mise en place pour garantir le respect des normes en matière de responsabilité, de transparence et de déontologie. La création d'un organisme indépendant qui s'engage en faveur des normes mondiales en matière de protection de la vie privée[124], possède des mécanismes d'application des normes et opère indépendamment du pouvoir exécutif permettrait d'atténuer ces problèmes.

Une régulation conjointe par le gouvernement, les médias sociaux et la société civile garantirait que les intérêts des utilisateurs sont au cœur des efforts de régulation. Les politiques gouvernementales autorisant la création dudit organisme de réglementation le protègeraient également contre la fluctuation des intérêts politiques et garantiraient la faisabilité de ses mécanismes d'application des normes. L'obligation pour les plateformes de respecter les mécanismes assurerait

120 Par exemple, en Allemagne, la loi NetzDG de 2017 sur les réseaux sociaux impose des sanctions pouvant aller jusqu'à 50 millions d'euros aux médias sociaux ne supprimant pas les contenus illégaux dans les 24 heures. Voir : http://wp.ceps.eu/wp-content/uploads/2018/11/RR%20No2018-09_Germany's%20NetzDG.pdf

121 Elhai, W. (2020), « Regulating Digital Harm Across Borders: Exploring a Content Platform Commission », *SMSociety'20: International Conference on Social Media and Society*, https://doi.org/10.1145/3400806.3400832, p. 223–4.

122 Voir Matsakis, L. (2 mars 2018), « YouTube Doesn't Know Where Its Own Line Is », *Wired*. Disponible à l'adresse : https://www.wired.com/story/youtube-content-moderation-inconsistent/

123 Ibid.

124 Par exemple, la Global Network Initiative. Voir https://globalnetworkinitiative.org/

une application plus équitable et juste des efforts de modération. Le droit à la vie privée, la liberté d'expression et la responsabilité doivent constituer les valeurs de base de l'organisme et étayer sa gouvernance.

Recommandation 2 : Prendre des mesures pour lutter contre les biais algorithmiques dans le développement de logiciels au point de conception

Les politiques et pratiques des plateformes en ligne en matière de modération et de conservation de contenu prévoient rarement une participation ou une responsabilité publiques[125]. Les algorithmes, souvent développés « à huis clos » dans le secteur, exercent une influence considérable sur l'expérience en ligne de milliards d'utilisateurs, alors qu'ils peuvent être extrêmement biaisés. Les algorithmes « apprennent à partir de certaines images, souvent choisies par les ingénieurs », blancs et de sexe masculin en grande majorité et de façon disproportionnée[126]. Ces biais ont entraîné de graves problèmes dans la vie réelle, tels que des logiciels identifiant les prévenus noirs comme plus à même de récidiver[127] ou l'application Google Photos confondant les photos d'un utilisateur noir avec des gorilles[128].

Sur le plan de la LEV, ces biais algorithmiques signifient que les contenus extrémistes non occidentaux sont peu reconnus et peu modérés. Puisque les plus grandes sociétés technologiques du monde ont leur siège en Occident, « les ingénieurs et dirigeants responsables de la conception des produits technologiques manquent peut-être de connaissances sur les moteurs de la violence et de la discrimination dans les cultures étrangères »[129]. La situation au Myanmar illustre les enjeux lorsque la modération et la suppression de contenus non occidentaux sont sous-développées. Les discours haineux en langue birmane publiés en ligne contre la communauté rohingya ont provoqué des violences à grande échelle. Pourtant, l'escalade de la haine raciste à l'encontre de cette minorité a eu lieu de manière quasi incontrôlée, Facebook n'ayant employé que deux examinateurs de langue birmane[130].

Pour lutter contre ces biais, les plateformes technologiques peuvent faire appel à l'expertise de la société civile et du monde de la recherche pour éclairer la phase de développement de logiciels. Elles devraient mener des audits approfondis et réguliers de leur utilisation des algorithmes, dont les résultats devraient être rendus publics afin de renforcer la responsabilité, la transparence et la confiance du public[131].

---

125 « Tackling the Information Crisis: A Policy Framework for Media System Resilience », The Report of the LSE Commission on Truth Trust & Democracy, p. 18. Disponible à l'adresse : https://www.lse.ac.uk/media-and-communications/assets/documents/research/T3-Report-Tackling-the-Information-Crisis-v6.pdf

126 Crawford, K. (25 juin 2016), « Artificial Intelligence's White Guy Problem », *New York Times. Disponible à l'adresse (accès payant) :* https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html

127 Angwin, J. *et al.* (23 mai 2016), « Machine Bias », *ProPublica*. Disponible à l'adresse : https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

128 Nieva, R. (1er juillet 2015), « Google apologizes for algorithms mistakenly calling black people "gorillas" », *CNET*. Disponible à l'adresse : https://www.cnet.com/news/google-apologizes-for-algorithm-mistakenly-calling-black-people-gorillas/

129 Elhai, p. 221.

130 Stecklow, S. (2018), « Special Report: Why Facebook is losing the war on hate speech in Myanmar », *Reuters*. Disponible à l'adresse : https://www.reuters.com/article/us-myanmar-facebook-hate-specialreport/special-report-why-facebook-is-losing-the-war-on-hate-speech-in-myanmar-idUSKBN1L01JY

131 Turner Lee, N. (2019), « Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms », Brookings Institute. Disponible à l'adresse : https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

Le développement des capacités en matière de modération de contenu – sur les plans linguistique et géographique pour les équipes humaines, et sur le plan du développement d'outils d'IA non-occidentaux – est un investissement coûteux pour les sociétés technologiques et les médias sociaux. Un plaidoyer énergique pour cette expansion indispensable peut offrir au secteur des technologies l'opportunité de se positionner comme des leaders en matière de stratégies de modération et de suppression de contenu.

Recommandation 3 : Assurer le financement, de la part des parties prenantes et initiatives nationales et multinationales, de publications qui présentent de façon claire et accessible la technologie, les enjeux et les opportunités liés à l'intelligence artificielle

Comme le fait remarquer l'auteure de *Intelligence artificielle et lutte contre l'extrémisme violent : rapport introductif*, l'IA fait grand bruit dans les médias. Nombreux sont les décideurs politiques qui ne comprennent pas bien de quoi il retourne, et les possibilités qu'offrent ces technologies. De plus, « les discussions parlementaires et auditions en commission au Royaume-Uni après le scandale Cambridge Analytica en 2018, au Congrès américain et au Parlement européen, ont mis au jour les niveaux extrêmement faibles d'éducation aux médias et de compréhension de la part des parlementaires et décideurs politiques de haut rang »[132].

Une mauvaise compréhension de l'environnement numérique et médiatique, ainsi que des capacités et limites de l'IA, peut entraîner des décisions politiques sous-optimales. Les décideurs politiques doivent être éduqués dans ce domaine, afin de prendre des décisions stratégiques éclairées, équilibrées et fondées sur des données factuelles.

Les initiatives nationales et multinationales doivent axer leurs efforts sur la production et la publication régulières d'orientations sur ce qu'est l'IA, ainsi que sur ses enjeux et opportunités dans le monde de la politique. La société civile et les universitaires spécialisés dans les préjudices en ligne et hors ligne dans des contextes particuliers devraient également produire des rapports introductifs réactifs sur les menaces émergentes et futures. Par exemple, les experts ayant des connaissances sur une élection controversée à venir dans un contexte non-occidental pourraient produire une note d'information destinée aux décideurs politiques sur le contexte et les catalyseurs des contenus préjudiciables, et expliquer comment ceux-ci pourraient donner lieu à de véritables préjudices dans la vie réelle.

Ces orientations et notes d'information devraient être rédigées de façon claire et accessible, et éviter le jargon technique ou sensationnaliste relatif à l'IA. Les décideurs politiques et non-spécialistes seraient alors en mesure de contribuer au discours public sur l'IA et la LEV.

132 « Tackling the Information Crisis: A Policy Framework for Media System Resilience », The Report of the LSE Commission on Truth Trust & Democracy, p. 38. Disponible à l'adresse : https://www.lse.ac.uk/media-and-communications/assets/documents/research/T3-Report-Tackling-the-Information-Crisis-v6.pdf

Global Network
on Extremism & Technology

COORDONNÉES

Pour toute question, demande d'information et demande de copies supplémentaires du présent rapport, contacter :

ICSR
King's College London
Strand
Londres WC2R 2LS
Royaume-Uni

T. **+44 20 7848 2098**
E. **mail@gnet-research.org**

Twitter : **@GNET_research**

Ce rapport peut, comme toutes les autres publications du GNET, être téléchargé gratuitement à partir du site Internet du GNET : www.gnet-research.org.