IRSEM
INSTITUT DE RECHERCHE STRATÉGIQUE DE L'ÉCOLE MILITAIRE

www.irsem.fr

Brève stratégique - 62

20 juillet 2023

Les guerres de l'information à l'heure de l'intelligence artificielle

DUSAN BOZALKA

L'intelligence artificielle (IA) favorise non seulement les capacités offensives d'acteurs malveillants sur le terrain numérique, mais elle fragilise aussi davantage les capacités défensives de nos sociétés face aux manipulations de l'information. Réalisée sur la base de discussions menées à la conférence annuelle du Centre d'excellence pour la communication stratégique de l'OTAN à Riga (Lettonie), cette brève stratégique se propose de synthétiser les dangers que présentent les outils utilisant cette technologie.

Si la démocratisation de l'Internet s'est initialement accompagnée d'une vision émancipatrice, nombreuses sont aujourd'hui les inquiétudes relatives à l'incursion croissante de l'IA dans nos sociétés. Force est de constater que **cette technologie est amenée à s'installer de manière durable dans notre quotidien**, comme le signale la mise à disposition d'outils en libre accès tels que des générateurs d'images et des agents conversationnels. Les prototypes les plus notoires de cette classe d'outils sont les deux dernières versions des transformateurs génératifs pré-entraînés (*Generative Pre-trained Transformer*) logés dans l'agent conversationnel développé par l'entreprise américaine OpenAI, ChatGPT. Grâce à une architecture complexe dite « en réseau neuronal », cette technologie exploite un modèle consacré au traitement d'une séquence de mots afin de prédire la meilleure réponse possible selon les données accessibles. Les outils générateurs d'images, tels que celui de *Midjourney*, utilisent quant à eux un modèle de diffusion, où du bruit aléatoire est progressivement inversé pour créer des images de haute qualité sur la base des instructions formulées par l'utilisateur. Ce faisant, ces outils demeurent **strictement façonnés par la main de l'humain**, nuançant le contenu des scénarios apocalyptiques liés à une automatisation complète de l'IA. Ils représentent toutefois un facteur d'intensification des ambitions poursuivies par les acteurs qui la mobilisent, ce qui inclut de multiples usages malveillants déjà observables sur le terrain numérique. Ce constat pousse de manière croissante les institutions internationales et leurs États membres à s'interroger sur les effets néfastes de l'IA, *a fortiori* dans un contexte géopolitique marqué par une recrudescence des guerres de l'information. Un des exemples les plus récents en est la conférence annuelle du Centre d'excellence pour la communication stratégique de l'OTAN, organisée en juin 2023 à Riga. Réalisée à la lumière des échanges entre chercheurs et militaires de haut rang, cette brève stratégique se propose d'énoncer deux enseignements relatifs aux menaces informationnelles que pose cette technologie.

L'IA incarne d'abord un **avantage pour les capacités offensives des États et des acteurs privés malveillants engagés dans la lutte informationnelle**. Ceci s'explique par la mobilisation croissante des outils issus de cette technologie en vue de massifier la création de fausses informations selon les requêtes formulées, et ainsi **contribuer à intensifier le brouillard informationnel**. Comme le montre une récente enquête de *NewsGuard*, **les réponses proposées par les modèles de ChatGPT en font régulièrement un relais de la propagande des gouvernements chinois et russe**. Une fois mises à contribution par des acteurs étatiques autoritaires, les réponses fournies servent à appuyer les récits de propagande et ceux véhiculés au travers d'opérations de désinformation. Bien que des garde-fous existent afin d'empêcher ces outils d'accéder à des requêtes considérées comme malveillantes, **la modération de ces dernières présente certaines failles**. Il est possible de débrider (*jailbreak*) ChatGPT, c'est-à-dire contourner les limitations mises en place par OpenAI, à la suite d'une requête précédée par une mise en situation

hypothétique ou introduite sous la forme d'un jeu de rôle (*prompt hacking*). En conséquence, certains spécialistes mettent en garde contre **la diminution des coûts entraînée par l'utilisation de l'IA dans les manipulations de l'information**. Ces outils sont en effet capables de **proposer des récits adaptés à des contextes culturels et linguistiques précis**, ciblant des groupes sociologiques variés et destinés à assurer une meilleure réception. Il est alors possible de produire des récits de propagande de haute qualité pour un coût relativement bas, l'IA remplaçant des agents sensibilisés aux enjeux nationaux. Apte à coder dans différents langages informatiques, ChatGPT facilite également la création de sites en ligne et de réseaux de comptes automatiques (*botnets*) destinés à amplifier artificiellement (*astroturfing*) les récits formulés. Au-delà des acteurs étatiques, cette tendance risque de multiplier les actes malveillants en provenance d'acteurs privés, et ce de deux manières différentes. La première se rapporte à **une nette augmentation de la création de *botnets* disponibles sur le *dark web* et actifs sur les réseaux sociaux**, au moment même où des plateformes telles que Twitter pâtissent des licenciements massifs de leurs modérateurs. La seconde concerne **la massification des opérations de cybercriminalité, telles que l'hameçonnage**, si bien que ChatGPT propose la formulation rapide de courriels frauduleux à même de contourner les filtres anti-spam et dotés d'un niveau de réalisme convaincant.

De ce premier point découlent des **limitations systémiques propres à la posture défensive des démocraties**, principalement ancrées dans des raisons techniques et économiques. **Les outils issus de l'IA ne sont pas en mesure de combattre les manipulations de l'information**, puisqu'ils sont incapables de les détecter. À ce jour, seules certaines **erreurs récurrentes** permettent d'identifier la marque d'une production artificielle, telles que des mains déformées ou des fonds flous lorsqu'il s'agit de générateurs d'images. Du côté des agents conversationnels, ce phénomène se matérialise à travers des tournures redondantes, la présence de statistiques improbables au sein du corps de texte, et la publication de messages d'erreur par des comptes automatiques (*bots*) sur les plateformes numériques. En matière de défense, l'utilité de l'IA réside *in fine* dans l'optimisation du temps de travail dont disposent les spécialistes des manipulations informationnelles. Ils peuvent alors se consacrer pleinement à l'identification de ces dernières et automatiser leur inventorisation. Si les modèles des intelligences artificielles sont accessibles à tous en libre accès, **leur développement et leur entraînement requièrent un investissement conséquent de plusieurs centaines de millions d'euros, résultant de la préparation d'une base de données exploitable et l'utilisation de serveurs informatiques puissants**. Cette limitation révèle une seconde difficulté d'ordre économique. De manière similaire aux plateformes numériques, **les données des échanges entre utilisateurs et les outils issus de l'IA restent la chasse gardée d'un oligopole d'entreprises, qu'il est impossible de réquisitionner comme c'est le cas dans les régimes autoritaires**. Ces lacunes techniques et financières empêchent les décideurs politiques et la communauté scientifique, faute de transparence et de moyens, de mieux comprendre les menaces posées par l'IA. Encore faut-il ajouter **les enjeux géopolitiques que risquent de provoquer les efforts de régulation de cette technologie**. Si les appels à la création de centres internationaux chargés de surveiller les évolutions liées à l'IA paraissent louables, ce processus reste fastidieux et compromis par la *Realpolitik* des États.

Les dangers informationnels que cette technologie représente pour la résilience des sociétés occidentales sont tangibles, mais **plusieurs réponses peuvent en mitiger les effets**. L'une de celles-ci implique **une augmentation conséquente des investissements consacrés à la recherche académique**. Davantage de travaux permettraient en effet d'améliorer la reconnaissance de messages formulés par des outils d'IA, notamment grâce à un modèle reposant sur une alliance de la stylométrie, soit la reconnaissance de tournures linguistiques redondantes, et de l'apprentissage machine. Une autre réponse consisterait en la mise à disposition des données récoltées par les entreprises privées, ce qui multiplierait les efforts consacrés à une utilisation défensive de leurs outils. Une récente étude souligne à ce titre le potentiel bénéfique de ceux-ci, dont ChatGPT, afin d'améliorer l'efficacité et la rapidité des processus de vérification des faits. Alors que l'IA menace la cohésion de nos démocraties dans leur ensemble, ces réponses nécessitent de **repenser leurs cadres juridiques**. Des règles claires sont en effet nécessaires afin d'engager la responsabilité des entreprises lorsque leurs outils sont employés dans des opérations de désinformation. **Le consensus académique semble quant à lui plaider en faveur d'une adaptation rapide des cadres juridiques nationaux, voire européens, aux dispositions applicables à l'IA**, telles que celles en matière de droits d'auteurs ou de la protection des données des utilisateurs. Des dispositions préventives peuvent parallèlement être adoptées face aux menaces posées par une utilisation autoritaire de l'intelligence artificielle, à l'image de la décision des États-Unis de contrôler l'exportation vers la Chine de processeurs. À cela s'ajoute un nécessaire **élargissement des méthodes utilisées par les acteurs gouvernementaux dans leur communication stratégique**, ce que soutiennent les effets positifs d'une exposition préventive à de la désinformation (*pre-bunking*). ■

Dusan Bozalka *est doctorant résident à l'IRSEM et rattaché au Centre d'analyse et de recherche interdisciplinaire sur les médias à l'Université Paris-Panthéon-Assas.*

Contact : dusan.bozalka@irsem.fr