UQÀM | CHAIRE RAOUL-DANDURAND EN ÉTUDES STRATÉGIQUES ET DIPLOMATIQUES

CHRONIQUES DES NOUVELLES CONFLICTUALITÉS

@ Image du pape générée par un utilisateur Reddit "u/trippy_art_special" sur Midjourney

Une image vaut mille mots : les périls de l'IA générative

Par Danny Gagné

La mise en service de ChatGPT en novembre 2022 a créé une onde de choc. Allions-nous être en mesure de départager le vrai du faux avec une intelligence artificielle (IA) de plus en plus performante ? Près d'un an plus tard, et au regard de la multiplication des programmes d'IA générative, ces craintes étaient-elles fondées ?

Le 22 mai dernier, une image insolite est publiée par le compte Twitter de RT, un média d'État russe. On peut y apercevoir une explosion à proximité du Pentagone, le siège du département américain de la Défense. La publication devient rapidement virale et va même jusqu'à affecter les marchés boursiers : le S&P500 chute de 0,3 % et les prix des bons du Trésor américain et de l'or augmentent au même moment. Le problème ? Cette image est en fait une pure invention. Elle a été créée grâce à un programme dit « text to image », c'est-à-dire un logiciel d'intelligence artificielle générative permettant de créer une image *ex nihilo* sur la base de mots clés entrés par un utilisateur.

La multiplication de ces programmes et la compétition entre les plateformes proposant de tels services, comme DALL-E 2 et Midjourney, permettent la création de fausses images de plus en plus convaincantes. Outre le cas du Pentagone, d'autres hypertrucages ont retenu l'attention récemment. En mars 2023, une image du Pape François 1er arborant un manteau Balenciaga est elle aussi devenue virale. Si le cliché a pu faire rire, il présentait néanmoins le pape accoutré d'un manteau valant des milliers de dollars — de quoi nourrir les critiques quant à l'opulence dans laquelle vivrait le souverain pontife. Quelques jours plus tôt, c'était la photo d'un Donald Trump résistant à son arrestation et aux prises avec de nombreux policiers qui enflammait la toile.

Ces deux images ont quelque chose en commun : elles ont été créées par de simples individus. Celle du pape est l'œuvre de Pablo Xavier, un travailleur de la construction de 31 ans de la région de Chicago qui s'amusait sur Midjourney. Celle de Trump est le fruit du travail d'Eliot Higgins, journaliste et fondateur du site d'investigation Bellingcat, qui tuait le temps en attendant la véritable arrestation de l'ancien président par les autorités new-yorkaises.

Ce phénomène a de quoi inquiéter. En effet, si de simples citoyens sont capables de générer autant de clics et de partages sur les médias sociaux, on peut se demander ce que des acteurs avec de mauvaises intentions et plus de ressources pourraient en faire.

La Chine passe à l'offensive

À l'aube de la prochaine élection présidentielle aux États-Unis, la Chine aurait déjà commencé son travail de sape contre la démocratie américaine. Si l'on imagine généralement la Russie mener ce genre d'activité, la Chine profite des développements de l'IA générative pour monter son jeu d'un cran. Depuis le début de l'année 2023, des images créées par IA ont été largement repartagées par de faux comptes sur différents médias sociaux, et des analystes de Microsoft attribuent cette campagne de propagande à des influenceurs chinois, possiblement à la solde de Pékin. Sans dévoiler de chiffres officiels, Microsoft assure que cette campagne a généré plus d'engagements de la part de véritables utilisateurs que lors de tentatives précédentes.

Parmi les fausses images relayées par ces influenceurs chinois, on retrouve, entre autres, la statue de la Liberté affublée d'une mitraillette ou encore des messages en soutien au mouvement *Black Lives Matter* accompagnés d'images de violence policière. Les publics cibles sont généralement les mêmes que la Russie avait visé en s'ingérant dans la campagne présidentielle de 2016, soit les franges de la population américaine les plus politisées (à droite comme à gauche), plus susceptibles d'être affectées par ces campagnes d'influence. La stratégie est assez claire : exploiter les divisions préexistantes au sein de la société américaine pour, parfois même, inciter les gens à sortir dans la rue.

Rapport de force affecté

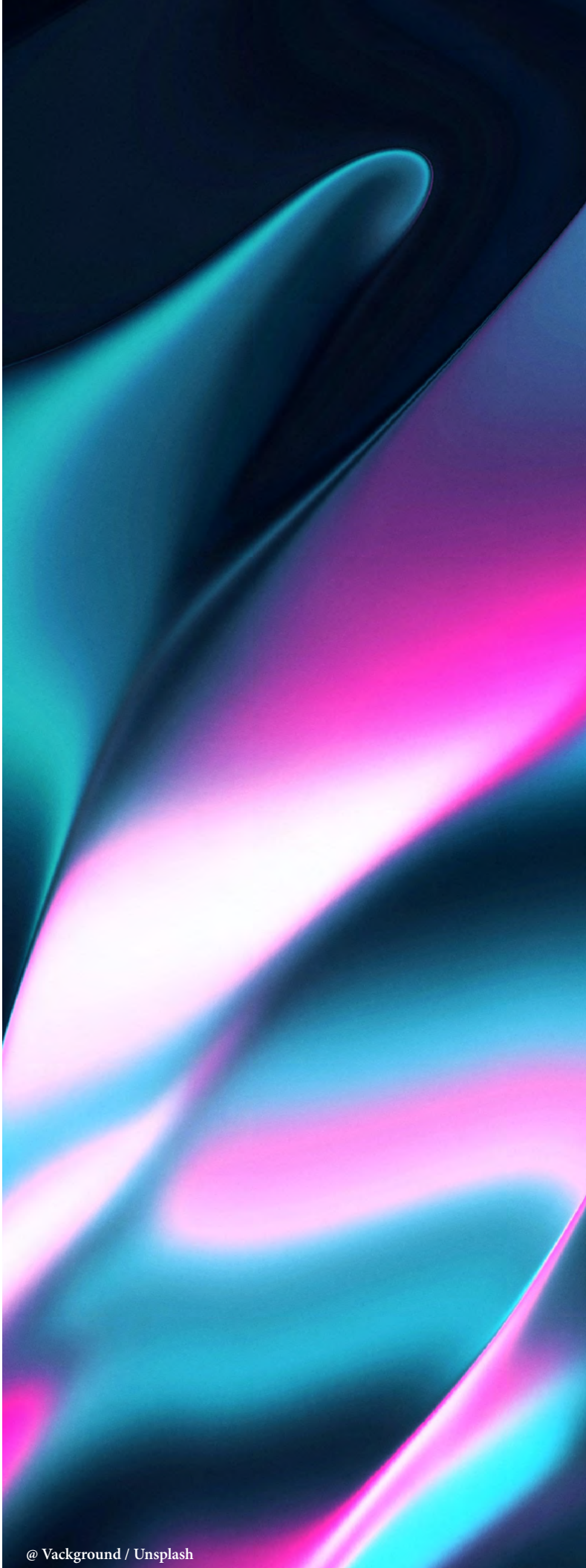
Selon une étude de la *School of Interactive Computing* du *Georgia Institute of Technology*, on observe que l'IA générative est justement plus habile que jamais pour cibler les sensibilités des internautes. Les programmes d'intelligence artificielle sont en mesure de saisir avec précision leurs préférences— en analysant leurs publications et les messages qu'ils partagent par exemple—, ce qui permet de créer des messages à la pièce, adaptés aux utilisateurs visés. Une même campagne d'influence peut ainsi se décliner en plusieurs versions légèrement différentes, mais en poursuivant un même objectif.

Si l'IA générative fait des progrès fulgurants, il est, pour l'heure, toujours possible de trouver certaines failles. Dans le cas de l'image de la statue de la Liberté, on remarque par exemple qu'une de ses mains possède six doigts. Trump semble quant à lui avoir trois jambes sur certains clichés de sa fausse arrestation. Néanmoins, puisqu'ils sont produits avec facilité et qu'ils circulent abondamment et rapidement, ces clichés placent parfois leurs victimes sur la défensive, car celles-ci doivent consacrer de plus en plus de temps et de ressources pour démontrer qu'ils sont faux.

De plus, les outils d'IA générative semblent affecter le rapport de force entre des médias traditionnels, aux prises avec une crise de confiance du grand public, et des médias « alternatifs » ou sensationnalistes qui relaient de telles images et amplifient ainsi les discours situés aux extrêmes pour s'assurer de la fidélité d'un certain auditoire.

Comment s'adapter ?

Le président américain Joe Biden est en discussion avec les géants de l'IA générative, dont OpenAI, créateur de ChatGPT. Au cœur des discussions, il



@ Vackground / Unsplash

est question de l'obligation d'inclure un filigrane sur les fausses images afin de freiner leur contagion et la désinformation. Or, les adversaires du président ont saisi tout le potentiel de ces outils et ont déjà commencé à s'en servir contre lui. En avril, une publicité électorale du Parti républicain, entièrement créée grâce à l'IA, illustrait une vision sombre de l'avenir des États-Unis si le président démocrate est réélu : raz-de-marée de migrants à la frontière sud, Troisième Guerre mondiale imminente et soldats armés patrouillant dans les rues d'une Amérique en ruine. En petit caractère, à peine visible sur la publicité, il était écrit : « Built entirely with AI imagery ».

Voyant le potentiel déstabilisateur de ces trucages, la plateforme Midjourney a commencé à restreindre l'utilisation d'images de certaines figures publiques telles que Xi Jinping. D'autres initiatives sont également encourageantes, comme l'alliance de CBC/Radio-Canada, la BBC, le New York Times et Microsoft dans le cadre du *Project Origin*. Cette collaboration vise à permettre aux organes de presse de certifier la provenance de leurs informations ainsi que la véracité des images et du contenu utilisés dans leurs publications.

Un problème de taille persiste néanmoins. Les internautes qui ont été convaincus par ces campagnes sont souvent insensibles au « fact-checking », leurs opinions étant justement fondées sur une méfiance des médias traditionnels, considérés comme étant au service des élites. Les velléités de régulation des contenus, quant à elles, s'apparentent à une boîte de Pandore. En effet, accusées par certains d'être orwelliennes, ces règlementations risquent d'accentuer davantage la dynamique de polarisation. En témoignent les débats tenus à Ottawa en 2022 quant à la définition du concept de « contenus préjudiciables ». Ainsi, s'il est désormais possible de générer automatiquement de la désinformation, produire des solutions viables pour y faire face semble un processus beaucoup plus laborieux.

Danny Gagné est chercheur à l'Observatoire des conflits multidimensionnels de la Chaire Raoul-Dandurand.

Pour en savoir plus sur la Chaire Raoul-Dandurand et ses travaux : https://dandurand.uqam.ca/

@ Tianyi Ma / Unsplash