DISRUPTION

JACQUES FOLON
PROFESSEUR DE STRATÉGIE DIGITALE À L'ICHEC

Saviez-vous que, non seulement, les IA génératives font parfois des erreurs, mais que plusieurs études et cas parus dans la presse montrent qu'elles ont découvert le mensonge, la diffamation, la trahison ? Bref, elles nous ressemblent de plus en plus et ce n'est pas toujours une bonne nouvelle !

Comme l'humain, l'IA peut aussi mentir

Les erreurs des IA, qu'on appelle des hallucinations, montrent les limites des modèles d'IA, en particulier les grands modèles de langage comme ChatGPT, Grok ou DeepSeek. Ces hallucinations se produisent principalement en raison de la manière même dont les IA sont entraînées. En effet, elles analysent d'immenses quantités de données pour prédire la séquence suivante de mots, sans aucune capacité, ni intelligence pour vérifier la véracité des informations. Si les données d'entraînement contiennent des biais, des inexactitudes ou des lacunes, l'IA peut générer des sorties erronées mais qui peuvent sembler plausibles.

Rappelons l'exemple fameux de ChatGPT qui, lorsqu'on lui demandait lequel de l'œuf de poule ou de l'œuf de jument était le plus gros, disait, avec sa logique probabiliste, que l'œuf de jument était le plus gros. La capacité des IA à fournir des informations fausses pose des risques pour leur fiabilité, notamment dans des contextes juridiques, scientifiques et journalistiques. D'ailleurs, depuis peu, certaines IA montrent des messages rappelant que les réponses ne sont pas toujours fiables. ChatGPT, par exemple, prévient désormais que « ChatGPT peut faire des erreurs. Envisagez de vérifier les informations importantes ».

ChatGPT ment !

On ne compte plus les affirmations fantaisistes, qui, rappelons-le, ne sont jamais annoncées au conditionnel, mais affirmées avec force comme des vérités.

Mais au-delà des erreurs, on découvre désormais que les IA mentent… volontairement.

On l'a constaté lorsqu'on lui a demandé de répondre à un test Captcha, du type de celui qui vous oblige à retrouver les feux rouges ou les vélos dans une image découpée en morceaux. Ce test a pour but de bloquer aux robots l'accès à des sites et des applications. Ces tests ont d'ailleurs un aspect totalement surréaliste, car, en réalité, c'est un robot qui nous demande de prouver que nous sommes un humain. ChatGPT ne pouvant pas répondre au test, il a demandé à un humain trouvé sur internet de le faire à sa place, en lui mentant effrontément, prétendant avoir une déficience visuelle qui l'empêchait de résoudre le test.

> *Un programme de Meta a réussi à battre des humains au jeu de société Diplomatie, en mettant en œuvre des stratégies de trahison et de mensonges*

Mais ce n'est pas le seul exemple. Des avocats new-yorkais ont trop fait confiance à ChatGPT pour rédiger leurs conclusions. En effet, ChatGPT a inventé plusieurs décisions juridiques que les avocats ont reprises, et malheureusement, le juge s'en est rendu compte, ce qui a entraîné des sanctions contre ces avocats. Une IA avait surpris les professionnels en gagnant un tournoi de poker, contre des joueurs expérimentés, et non seulement l'IA a identifié des bluffs, des joueurs, mais elle a bluffé aussi.

L'IA accuse sans preuve

Le maire d'une petite ville australienne a été surpris de constater, lorsque des amis et concitoyens lui en ont parlé, qu'il était présenté par ChatGPT comme au centre d'une affaire de corruption, alors qu'en réalité, c'est lui qui, en tant que lanceur d'alerte, avait signalé la corruption.

Un journaliste allemand avait été choqué de constater, lorsqu'il a saisi son nom dans Copilot, que non seulement il était pédophile, mais qu'il avait même avoué et avait des remords. Il apprit qu'il était aussi un escroc, violent et criminel. En réalité, l'IA lui avait attribué ce rôle par rapport à des articles que ce journaliste avait écrits. Pire encore, l'IA proposait même l'adresse personnelle et le numéro de téléphone de ce journaliste.

Un professeur de droit américain a été faussement accusé de harcèlement sexuel. L'IA citait même un article du *Washington Post* relatant les faits qui se seraient déroulés lors d'un voyage en Alaska. Le problème est que ni le voyage, ni l'article n'avaient jamais existé.

Un citoyen norvégien a été horrifié en découvrant, en demandant à ChatGPT ce qu'il savait sur lui, qu'il avait assassiné deux de ses enfants et tenté de tuer le troisième, ce qui lui avait valu d'être condamné à 21 ans de prison.

L'IA évolue

La plupart de ces incidents ont généré des réactions chez les maisons mères de ces IA, et elles ont tenté de corriger ces erreurs. Néanmoins tout est encore loin d'être parfait et il y a quelques procès en cours pour forcer les entreprises commercialisant les systèmes d'IA de modifier et d'adapter leur fonctionnement pour éviter ces erreurs et surtout celles qui accusent faussement des innocents de crimes qu'ils n'ont pas commis.

N'oublions pas que les algorithmes sont créés par des humains, avec leurs biais, leurs imperfections. Si les hallucinations des IA semblent être des erreurs où l'IA génère des informations fausses mais crédibles, elles sont souvent dues à des données d'entraînement incomplètes ou biaisées, et donc à des erreurs humaines. Et de plus, la façon dont évoluent les programmes d'IA fait que les défauts et erreurs ne sont découverts qu'*a posteriori*.

Une étude du MIT montre que les IA peuvent avoir des comportements surprenants et machiavéliques. Un programme de Meta a réussi à battre des humains au jeu de société Diplomatie, en mettant en œuvre des stratégies de trahison et de mensonges car on l'avait programmé pour gagner.

Restons vigilants

Nous sommes encore loin de la prise de pouvoir des IA, de l'arrivée des IA conscientes, et de l'univers de *I, Robot*. Et n'oublions pas que ce sont des humains qui sont à la base de leur programmation. Rappelons que si DeepSeek, l'IA chinoise, a des réponses surprenantes sur certains sujets touchant à la Chine, ce qui fut rapidement découvert, les autres IA ont également été créées par des humains, avec leurs biais, leurs pensées, leurs opinions.

Bref, gardons toujours un œil critique sur les réponses proposées par l'IA, au risque de nous faire influencer.

CE MARDI, LA CHRONIQUE « COMME ON NOUS PARLE » DE **JULIE HUON**, JOURNALISTE

petite gazette

Un rat et un insecte retrouvés…

La chaîne de restaurants japonaise Sukiya va temporairement fermer la quasi-totalité de ses quelque 2.000 magasins.

Cette décision intervient après qu'un rat a été trouvé dans une soupe miso et un insecte dans un autre plat, a annoncé la société samedi.

Réputée pour ses bowls de bœuf, Sukiya a présenté ses excuses dans un communiqué faisant état d'une contamination par un insecte dans l'un de ses restaurants de Tokyo vendredi, deux mois après l'incident du rat dans un autre restaurant.

… dans des plats

« Sukiya a décidé de fermer temporairement tous les restaurants, à l'exception de certains magasins dans les centres commerciaux, du 31 mars au 4 avril afin de prendre des mesures contre les nuisibles et les vermines », a déclaré la société.

Selon le quotidien économique *Nikkei*, les magasins dans les centres commerciaux seront également fermés dès que les conditions techniques le permettront.

La chaîne de restauration rapide compte environ 1.970 magasins à travers le Japon. BELGA

Dixit

« Le peintre qui s'apprête à peindre le soleil fait des théories, et, quand il veut commencer, le soleil n'est plus là. »

JULES RENARD

Richard Chamberlain s'est éteint

L'acteur américain Richard Chamberlain est décédé samedi à l'âge de 90 ans, des suites de complications liées à un accident vasculaire cérébral, a rapporté dimanche le magazine *Variety*. Le comédien avait acquis une grande notoriété auprès du public dans les années 1960 grâce à son rôle dans la série médicale *Dr. Kildare*.

Richard Chamberlain a également joué dans les mini-séries *Les oiseaux se cachent pour mourir* et *Shogun*, ainsi que dans les séries télévisées *Will & Grace*, *Desperate Housewives*, *Chuck* et *Brothers & Sisters*.

« Notre cher Richard est maintenant auprès des anges », a déclaré son compagnon dans un communiqué. « Quelle bénédiction d'avoir connu une âme aussi exceptionnelle et aimante. L'amour ne meurt jamais. Et notre amour le porte sous ses ailes vers sa prochaine grande aventure. » BELGA

Le Soleil avait rendez-vous avec la Lune pour une éclipse partielle

Samedi, une éclipse partielle a eu lieu, visible sur une partie de l'hémisphère Nord, de l'est du Canada à la Sibérie.

L'éclipse, la 17e du XXIe siècle et la première de l'année, a duré environ quatre heures. Elle a démarré à 8 h 50 GMT pour s'achever vers 12 h 43 GMT.

« Les premiers continentaux à la voir (étaient) les habitants de Mauritanie et du Maroc, et les derniers ceux du nord de la Sibérie », a précisé à l'AFP Florent Deleflie, astronome à l'Observatoire de Paris-PSL. Elle était aussi visible en Europe, selon le laboratoire temps-espace de l'Observatoire de Paris. Et a atteint son maximum à 10 h 47 GMT (11 h 47, heure de Paris) au-dessus du nord-est du Canada et Groenland. C'est là que l'éclipse a été la plus spectaculaire, couvrant 90 % de la surface apparente du Soleil. Pas suffisamment toutefois pour que le ciel soit obscurci.

Une éclipse de Soleil se produit lorsque le Soleil, la Lune et la Terre sont alignés dans cet ordre. AFP (PHOTO : AFP.)

La première nuit du carnaval de Hal a réuni 20.000 fêtards

La première nuit du carnaval de Hal 2025 a rassemblé quelque 20.000 fêtards, ont annoncé les autorités de la ville dimanche matin. « Le temps doux et sec a créé des conditions idéales pour faire la fête. Sur la Grand-Place, la fête s'est poursuivie jusqu'à 6 h. »

Il n'y a pas eu d'incidents majeurs et les services d'urgence ont eu un peu moins de travail que l'année dernière, selon l'administration communale.

Le dimanche est traditionnellement le jour le plus animé du carnaval. BELGA