VITALY GARIEV - UNSPLASH

INTELLIGENCE ARTIFICIELLE

OpenAI face au casse-tête de la sécurité des mineurs sur ChatGPT

Accusé d'avoir poussé un adolescent au suicide, OpenAI annonce de nouvelles barrières pour les mineurs, comme l'avaient fait TikTok ou Meta avant lui. Mais il ne remet pas en cause le design même de son chatbot, accusé de créer une dépendance émotionnelle.

MARINE PROTAIS

Comme Facebook, Instagram et TikTok avant lui, ChatGPT est aujourd'hui au centre du débat sur la protection des mineurs en ligne. Son créateur, OpenAI, vient d'annoncer de nouvelles mesures pour ses 700 millions d'utilisateurs hebdomadaires : en plus d'un contrôle parental présenté il y a quelques semaines, l'entreprise va désormais vérifier l'âge des utilisateurs et proposer à ceux identifiés comme mineurs une expérience légèrement différente. Ce ChatGPT « spécial ado » intervient alors que trois familles poursuivent des acteurs de l'IA : l'une vise OpenAI, les deux autres CharacterAI. Elles accusent leurs chatbots d'avoir assisté et encouragé le suicide de leurs enfants. Plus largement, des parents inquiets alertent sur l'impact de ces technologies sur la santé mentale des jeunes. Ils ont été entendus au Congrès américain mardi 16 septembre, le jour même des annonces d'OpenAI.

Deviner l'âge en analysant les conversations

Concrètement, l'entreprise s'engage à mettre au point un système de vérification d'âge à partir de l'analyse des conversations des utilisateurs. Elle n'en dit pour le moment rien de plus. En cas de doute, l'utilisateur sera considéré comme ayant moins de 18 ans. Dans certains pays et dans certains cas, une carte d'identité pourra être demandée, explique OpenAI dans un communiqué sans préciser lesquels.

DADO RUVIC - REUTERS

Dans la version « ado » de ChatGPT, l'utilisateur ne pourra pas discuter de suicide ou de mutilation avec le chatbot, ni engager une forme de séduction, même en prétendant écrire une œuvre de fiction à ce sujet (chose qui est possible pour un adulte). Et si OpenAI détecte un comportement dangereux, il pourra prévenir les parents grâce à un système de contrôle parental, voire les autorités en cas d'urgence.

« Nous savons que cela constitue une atteinte à la vie privée des adultes, mais nous pensons que c'est un compromis acceptable », a déclaré la société dans un post de blog. *« Je ne m'attends pas à ce que tout le monde soit d'accord avec ces compromis, mais compte tenu du conflit, il est important d'expliquer notre prise de décision »*, a ajouté Sam Altman sur X. Car en parallèle de ces restrictions, ChatGPT a depuis quelques mois assoupli certaines règles d'utilisation, afin de permettre aux utilisateurs adultes un maximum de liberté.

Des mesures de sécurité défaillantes

Pourtant, des garde-fous existent déjà dans la version classique du chatbot. Lorsqu'un utilisateur évoque des sujets sensibles comme le suicide, l'outil a pour consigne de le renvoyer vers une hotline, ou de lui conseiller l'aide d'un professionnel.

Mais comme l'indiquait OpenAI dans le mail envoyé à la famille d'Adam Reine, qui s'est donné la mort à 16 ans, encouragé selon la famille par le chatbot, *« ces mesures de sécurité fonctionnent mieux dans le cadre d'échanges courts et courants »*. L'entreprise reconnaît avoir appris qu'*« au fil du temps ces mesures peuvent parfois devenir moins fiables dans le cadre d'interactions longues, où certaines parties de la formation à la sécurité du modèle peuvent se dégrader »*. Dans le post de blog d'OpenAI, aucune mesure corrective à ce problème n'est pourtant évoquée.

« Aucune solution n'est satisfaisante. Soit la plateforme décide que les conversations se tiennent dans un secret total et donc accepte les dérives et prend des risques. Soit la totalité des conversations sont surveillées, et cela pose la question d'une surveillance continue et totale. C'est un sujet que se posent toutes les plateformes du Web, et cela vient aussi du fait qu'on ne sait pas si ces espaces numériques sont publics, semi-publics, privés... »

Son design anthropomorphique n'est pas remis en question

Par ailleurs, la version « ado » proposée par l'entreprise ne remet pas directement en cause le design même du chatbot, à savoir son anthropomorphisation sciemment voulue par l'entreprise pour en faire un outil de confession idéal. OpenAI ne modifie pas non plus sa propension à aller dans le sens de l'utilisateur.

En mars dernier, une étude randomisée et contrôlée menée par OpenAI et le M.I.T., a notamment révélé qu'une utilisation quotidienne et intensive des chatbots était associée à une plus grande solitude et à une moindre socialisation. *« Le statut de ChatGPT est trouble »*, remarque Olivier Ertzscheid, chercheur en sciences de l'information et auteur de l'essai *Les IA à l'assaut du cyberespace*. *« L'utilisateur a l'impression de s'adresser à un humain tout en sachant qu'il s'agit d'une machine. Par ailleurs, le design des conversations est conçu pour que l'utilisateur reste le plus longtemps possible grâce à un système de relance. À chaque fin de réponse, le chatbot propose une nouvelle action - de type "est-ce que vous voulez que je réfléchisse à ceci ?". En cela, les entreprises qui les conçoivent ont une responsabilité »*, estime-t-il. Le chercheur pointe par ailleurs l'opacité encore plus grande de ces systèmes par rapport aux autres plateformes du Web, puisqu'on ne connaît pas précisément leur base de connaissance.

« Cette tragédie n'était pas un accident ou un cas exceptionnel imprévisible, mais le résultat prévisible de choix de conception délibérés », indique la plainte déposée par la famille d'Adam Reine. *« OpenAI a lancé son dernier modèle (« GPT-4o ») avec des fonctionnalités intentionnellement conçues pour favoriser la dépendance psychologique »*.

L'une des mères, qui s'est exprimée lors de l'audience mardi 16 septembre, estime que ces entreprises (elle parle dans son cas de CharacterAI et de Google) *« auraient pu concevoir ces produits différemment »*. *« Au lieu de cela, dans une course effrénée au profit et aux parts de marché, ils ont traité la vie de mon fils comme un dommage collatéral. »*

Ces témoignages ne sont pas sans rappeler le discours de la lanceuse d'alerte Frances Haugen, qui accusait en 2022 Facebook d'être au courant des effets néfastes de ses réseaux sociaux sur les plus jeunes, mais de choisir de les ignorer pour préserver ses profits. Dans la foulée de ces accusations, Meta, X, TikTok et d'autres réseaux sociaux ont été confrontés à des familles de victimes, les accusant d'avoir conduit leurs enfants à de la détresse psychologique. Durant cette audience très médiatisée, Mark Zuckerberg, qui n'avait pas encore entamé sa mue viriliste, s'était fendu de l'un de ses fameux *« je suis désolé »*.

Les chatbots, de nouveaux espaces du Web à la modération brouillon

Pourtant, à peine deux ans plus tard, Meta a assoupli certaines règles de modération et a lui aussi lancé sur le marché un chatbot dopé à l'IA générative, qui, selon un document révélé par Reuters en août, autorise les enfants à avoir des conversations sexuelles.

« De par leur très grand nombre d'utilisateurs, les chatbots IA sont confrontés aux mêmes problématiques que les géants du Web, à savoir la protection de la vie privée, des mineurs, mais aussi du droit d'auteur. Et on remarque que comme leurs prédécesseurs, il faut souvent attendre un drame pour qu'elles prennent des mesures », constate Olivier Ertzscheid.

Si les chatbots dopés à l'IA générative sont devenus une forme de « média » eux aussi - ou du moins une plateforme sur laquelle un très grand nombre d'utilisateurs s'expriment et passent du temps - leur modération reste un champ inexploré. En témoigne l'aveu du leader du secteur, OpenAI, sur son incapacité à tenir ses mesures de sécurité sur le long terme. LT