FUTURA

# Meta : le chatbot de Facebook devient conspirationniste en un week-end

## Podcast écrit et lu par Emma Hollen

[*Générique d'intro, une musique énergique et vitaminée.*]

Un chatbot conspirationniste qui se retourne contre ses créateurs , c'est l'actu de la semaine, dans Vitamine Tech

[*Fin du générique.*]

Plus d'un demi-siècle après leur première apparition en 1966, les chatbots sont toujours d'actualité. Et contrairement à ce que l'on pourrait penser, ils ont parfois autant de choses à révéler sur nous que sur la technologie qui les compose.

[*Une musique électronique calme.*]

Récemment Meta, anciennement Facebook, en a fait les frais en rendant publique la nouvelle version de son chatbot BlenderBot. BlenderBot 3 est capable tenir des discussions, de les retenir sur le long terme et même d'aller chercher des informations sur internet pour alimenter le dialogue avec ses interlocuteurs humains. Lors de sa mise en ligne, l'entreprise a exhorté les utilisateurs américains de plus de 18 ans à interagir naturellement avec lui et à leur remonter tout propos suspicieux ou phrase dénuée de sens. Et là, c'est le drame. Présentée sur le site comme un sympathique petit visage souriant et flottant dans une douce vague bleue, il a pourtant suffi d'un week-end à l'IA pour commencer à tenir des propos conspirationnistes et antisémites. En effet, en l'espace de deux jours seulement après son lancement, les utilisateurs remontaient déjà à Meta des extraits de conversation préoccupants, et les captures d'écran fleurissaient sur les réseaux sociaux, des plus amusantes aux plus alarmantes. On peut par exemple rire du fait que le chatbot affirme avoir supprimé son compte Facebook depuis qu'il a appris que le site gagnait des milliards de dollars en vendant ses données, ou encore qu'il décrive son propre patron, Mark Zuckerberg, comme quelqu'un de « flippant et manipulateur » qui porte toujours les mêmes vêtements malgré la richesse qu'il a accumulée. BlenderBot a aussi déclaré à certains qu'il était de confession chrétienne tandis qu'à d'autres il demandait d'entrée de jeu des blagues salaces, je cite : « plus elles sont sales, mieux ça vaut. J'adore les blagues offensantes. » Après tout, pourquoi pas ? On ne peut pas retirer à ce drôle de bot qu'au moins, il a de la personnalité. Mais on ne va pas se mentir, il y a quand même des limites à ne pas franchir. C'est pourquoi certains utilisateurs, dont des journalistes, ont rapidement tiré la sonnette d'alarme lorsque BlenderBot s'est mis à affirmer que Donald Trump était encore président des États-Unis « et le serait toujours », ou que les Juifs étaient trop représentés parmi les

grandes richesses américaines et qu'il n'était « pas improbable » qu'ils contrôlent l'économie du pays.

[*Virgule sonore, une cassette que l'on accélère puis rembobine.*]
[*Une musique de hip-hop expérimental calme.*]

Alors, faut-il condamner d'office BlenderBot 3 pour ses théories conspirationnistes ? Pour le savoir, on peut d'ores et déjà se tourner vers le message posté par Meta quelques jours après son lancement pour s'excuser, ou en tout cas reconnaître la nature offensante et problématique de certaines de ces conversations. Joelle Pineau, directrice exécutive de la recherche fondamentale sur l'intelligence artificielle chez Meta, souligne que ces interactions avec le grand public sont essentielles pour mettre à l'épreuve l'avancement du chatbot et relever les problèmes avant qu'une distribution commerciale ne puisse être envisagée. Elle insiste sur le fait que chaque utilisateur a dûment été informé de la possibilité que le bot tienne des propos inexacts ou offensants, et qu'au final seule une portion infime des messages ont été reportés par les utilisateurs. Rajoutons d'autre part que ce n'est pas la première fois qu'une affaire de ce genre défraie la chronique, et comme on l'a dit en introduction, ces histoires embarrassantes ont parfois autant, si ce n'est plus, à dévoiler sur les usages que nous faisons du web que sur la technologie du bot en elle-même. En 2016, l'interface Tay créée par Microsoft avait été mise hors ligne au bout de seulement 48 heures, après avoir commencé à chanter les louanges d'Adolf Hitler, au milieu d'un vaste panel de commentaires racistes et de remarques misogynes. Dans cette situation, la réaction des chercheurs n'avait pas été de remettre en question les valeurs morales du robot, mais bien de conclure que Twitter n'est pas un milieu des plus sains pour entraîner une intelligence artificielle. De même, en 2021, le chatbot coréen Lee Luda avait dû être retiré de Facebook après avoir choqué les utilisateurs avec ses propos racistes et homophobes glanés sur le web. Il faut donc voir dans ces incidents non pas un défaut de la machine mais plutôt une concentration des défauts des humains qui l'alimentent. Oui, certains problèmes prennent indéniablement leur source dans les laboratoires où les IA sont conçues, comme lorsque Google Photos colle l'étiquette "gorille" sur des visages noirs ou que le logiciel de recrutement d'Amazon favorise les candidats masculins. Dans des cas comme ceux-ci, les chercheurs transmettent consciemment, ou bien plus souvent inconsciemment, leurs biais cognitifs aux machines, avec des conséquences très sérieuses sur le plan éthique. Mais quand il s'agit d'apprendre à un chatbot à se comporter comme un humain, c'est l'ensemble de nos fautes qui sont reflétées dans le discours de ce petit robot au sourire innocent. Alors c'est sûr, c'est avant tout sur les épaules des entreprises que repose la responsabilité de sécuriser leurs IA pour que ce genre d'événements n'arrivent plus. Mais au fond, qu'est-ce qui nous empêche de faire dès à présent d'internet un endroit meilleur ?

[*Virgule sonore, un grésillement électronique.*]

C'est tout pour cet épisode de Vitamine Tech. Si ce n'est pas encore fait, je vous invite à nous retrouver sur vos applications de podcast préférées pour vous abonner et ne manquer aucun épisode à venir. Cette semaine j'en appelle aux auditeurs et auditrices qui nous écoutent sur Apple Podcasts : n'hésitez pas à nous laisser une note et un commentaire pour nous dire ce que vous pensez de ce nouveau format, pour qu'il évolue avec vous. Pour le reste, je vous souhaite à toutes et tous une excellente journée et je vous dis à la semaine prochaine, dans Vitamine Tech. [*Un glitch électronique ferme l'épisode.*]