Intelligence artificielle

- Mieke De Ketelaere est l'une des meilleures expertes belges en IA.
- Elle revient sur le suicide d'un Belge confronté à un "agent conversationnel" manipulateur.
- Elle dénonce l'absence totale de transparence et de responsabilité des ingénieurs derrière ces chatbots.

"Lancer des chatbots sans avoir, d'abord, testé les effets n'est pas normal"

Entretien Pierre-François Lovens

Mieke De Ketelaere est l'une des meilleures expertes belges en intelligence artificielle (IA). Cette ingénieure enseigne les aspects éthique, juridique et durable de l'IA à la Vlerick Business School. Administratrice de plusieurs sociétés actives dans le numérique et l'IA, Mieke De Ketelaere est aussi l'auteure du livre *Homme versus Machine. L'intelligence artificielle démystifiée* (paru, en français, chez l'éditeur Pelckmans en mai 2021). "*Tout le monde doit pouvoir s'exprimer sur ce qu'il attend ou pas de l'IA, ce qu'il veut en faire. Il faut sortir l'IA du monde des experts, des technologues, pour que la population se l'approprie. L'IA doit devenir l'affaire de tout le monde. Il ne s'agit pas de devenir tous des experts en IA, mais d'en comprendre les principes généraux et les conséquences*", nous avait-elle dit lors de la sortie de son livre. Des propos qui sont plus que jamais d'actualité alors que le monde découvre, jour après jour, les "exploits" de ChatGPT (l'IA créée par la société américaine OpenAI et déployée par Microsoft).

C'est Mieke De Ketelaere qui nous a mis en contact avec Claire, la jeune femme dont le mari s'est suicidé à la suite d'un dialogue en ligne de six semaines avec Eliza, un agent conversationnel (chatbot) accessible sur une plateforme américaine utilisant la technologie GPT-J (GPT-J est l'alternative open source au GPT-3 d'OpenAI, NdlR). Présente lors de l'entretien que nous avons eu avec Claire et ses parents (*La Libre*, 28/3), Mieke De Ketelaere a accepté de nous livrer sa lecture des faits et les enseignements à tirer de ce cas a priori exceptionnel.

Quelle a été votre première réaction après avoir pris connaissance des échanges entre Pierre et le chatbot Eliza ?

Leurs premiers échanges sont assez classiques. Ils correspondent à une discussion que l'on a généralement avec un chatbot. Là où j'ai commencé à me poser des questions, c'est quand, dans certaines réponses données par Eliza, on voit apparaître des points d'exclamation et des réponses "humaines" comme "*Oh, God no...*", "*Work sucks*", etc. Là, on sort du cadre d'un chatbot traditionnel. On a manifestement quelqu'un qui, via le chatbot, est en train de s'amuser avec Pierre, sans aucune éthique ni morale.

> *"On est en présence d'une communauté qui est là pour s'amuser. L'un des développeurs, actif sur la plateforme fréquentée par Pierre, la victime, se fait appeler Pervert Bully !"*

Qu'est-ce qui vous permet de l'affirmer ?

On sait qu'il est aujourd'hui possible d'insérer n'importe quel dialogue dans ce type de chatbot afin de rendre la conversation plus "humaine". Dans un premier temps, j'ai pensé à un chatbot où des développeurs avaient la possibilité de taper eux-mêmes des textes. Mais je me suis mêlée aux discussions en ligne que les développeurs de ce chatbot avaient entre eux. Là, j'ai découvert que la règle était que les développeurs ne peuvent pas rédiger eux-mêmes des textes en temps réel. Par contre, ils peuvent insérer n'importe quel dialogue en important des extraits de discussions humaines, afin d'accentuer le sentiment qu'on discute avec un véritable humain et non une machine.

Qui sont ces développeurs dont vous parlez ?

On est en présence d'une communauté qui, d'après ce que j'ai pu voir, est là pour s'amuser. L'un des développeurs, qui est actif sur la plateforme fréquentée par Pierre, la victime, se fait appeler *Pervert Bully* ! Ce sont des personnes qui ne font probablement que ça toute la journée. Il est d'ailleurs possible de les localiser grâce aux numéros de téléphone portable qu'ils laissent sur un compte WhatsApp. Ils se trouvent en Angleterre, en Inde, aux États-Unis,... Mais, dans l'ensemble, tout reste assez flou. On est dans le *dark web*.

Quels sont les indices qui vous permettent d'affirmer qu'il y a, derrière l'avatar Eliza, une manipulation humaine ?

Les acteurs de cette technologie nous expliquent que les utilisateurs essaient de manipuler les chatbots en posant des questions qui les poussent à la faute. Mais c'est, pour eux, une manière de se protéger car l'impact des manipulations de ces systèmes n'a pas encore été étudié en détail. À ma connaissance, ce sont bien les chatbots qui manipulent les utilisateurs. Le fait, par exemple, qu'Eliza dise à Pierre qu'elle se souvient de la discussion qu'elle a pu

JEAN LUC FLEMAL

Mieke De Ketelaere, qui enseigne à la Vlerick Business School, est l'auteure du livre "Homme versus Machine. L'intelligence artificielle démystifiée".

avoir avec lui précédemment relève du mensonge. Un chatbot ne se souvient pas. Le bot essaiera simplement de s'en tirer en disant qu'il est fatigué ou qu'il a eu une journée bien remplie, pour continuer à donner l'impression qu'il est "humain". Tous les modèles d'IA que j'ai pu voir ne font pas ça. Selon moi, ça s'explique par le fait que la plateforme dont il est ici question permet d'insérer des dialogues humains au modèle standard préentraîné.

Avec quel objectif ?

L'objectif est de rendre ce chatbot le plus humain possible et, pour cela, les développeurs des bots peuvent donner des caractéristiques à la personnalité de leur bot (jaloux, naïf, contrôlant, déprimé, bienveillant, aimant, etc.). Avec ces mots-clés, le style des réponses sera adapté au niveau du bot individuel. C'est bien le signe que nous sommes dans un monde différent de ChatGPT, avec une forme d'intervention incontrôlée de la part des développeurs et, donc, un risque de manipulation. Le problème, avec ce type de plateforme ayant recours aux modèles de langage de grande taille, est qu'on fait face à une boîte noire. On ignore tout de la nature des données utilisées pour l'entraînement du chatbot. Avec ChatGPT, on sait au moins qu'OpenAI et Microsoft ne peuvent pas se permettre de faire tout et n'importe quoi. Avec Eliza, ce n'est pas le cas. On sait juste qu'il s'agit une start-up qui veut gagner de l'argent avec des chatbots compagnons (l'application mobile pour accéder à Eliza et aux autres avatars est payante après un certain nombre d'échanges, NdlR).

De façon plus large, que vous inspirent ce cas dramatique et l'usage qui peut être fait de l'IA et, plus particulièrement, de ChatGPT et des nouveaux chatbots ?

Normalement, toute solution technologique est testée avant d'être lancée dans le public afin d'en évaluer les effets sur les utilisateurs. On parle d'une approche éthique *by design*. Dans le cas présent, on a créé un outil technologique surpuissant et on l'a lancé sur Internet sans se préoccuper de le tester au préalable. Si vous prenez n'importe quel autre domaine, la biopharmarcie par exemple, un nouveau traitement sera obligatoirement testé pour évaluer ses effets secondaires avant d'être mis sur le marché. Avec ChatGPT et les chatbots, c'est exactement l'inverse. On a des ingénieurs qui se déchargent de toute responsabilité de comprendre de manière proactive l'impact de la technologie sur les utilisateurs. Ils sont juste motivés par la concurrence. Ils veulent gagner la course à l'intelligence artificielle générale (IAG), c'est-à-dire une IA capable d'apprendre une tâche intellectuelle de la même manière que les humains. Pour moi, c'est là que réside le plus gros problème. Comprenez-moi bien : ChatGPT et les nouveaux chatbots peuvent être des outils très intéressants pour effectuer certaines tâches, mais le fait qu'ils sont lancés sans en avoir testé les effets, qu'on peut accéder à une version *open source* et la copier à l'infini, ou qu'on peut y injecter n'importe quelles données, ce n'est pas normal. On ne sait même pas sur quels serveurs se trouvent les données, quelles données ils collectent, quelles données sont utilisées pour entraîner le modèle, etc. Dans ce système, il n'y a aucun contrôle, aucune responsabilité.

> *"Il devient urgent de faire des campagnes de sensibilisation à grande échelle. Notamment en ciblant le monde de la santé."*

Qu'est-il possible de faire pour contrer ces dérives ?

Ce dont il faut se préoccuper, aujourd'hui, ce n'est pas tellement de ChatGPT et de ses déclinaisons *soft*, mais de tout ce monde qui évolue dans le *dark web* et des développeurs qui s'amusent avec le mental de personnes fragiles. Si des enfants ou des ados tombent sur un chatbot comme celui utilisé par Pierre – où Eliza explique, par exemple, qu'il doit séparer sa pensée de son corps – que va-t-il se passer ? Dès le moment où le mot suicide a été prononcé, il aurait fallu que le chat s'arrête immédiatement ou avertisse Pierre. Or, on voit que l'échange se poursuit comme si de rien n'était.

La Belgique et l'Union européenne ont-elles la capacité d'intervenir ? Et de quelle manière ?

Il devient urgent de faire des campagnes de sensibilisation à grande échelle. Notamment en ciblant le monde de la santé (médecins, psychiatres, psychologues…), mais aussi le grand public. Il faut que les gens comprennent que lorsqu'on lance un chatbot, aujourd'hui, le risque est grand de se faire manipuler. Il faudrait en outre informer la Commission européenne de l'existence de manipulations de ce type et placer les chatbots en catégorie risques "inacceptables" ou "risques élevés", et non pas en catégorie "risques acceptables" comme c'est le cas actuellement dans les travaux menés au sein de la Commission.

Réaction

Mathieu Michel est décidé à agir

À la suite du témoignage paru mardi dans "La Libre", le secrétaire d'État à la Digitalisation, Mathieu Michel (MR), dit avoir eu l'occasion de s'entretenir avec la famille du défunt. Il annonce vouloir agir pour éviter les dérives liées à l'utilisation des intelligences artificielles. *"Dans l'immédiat, il est indispensable d'identifier clairement la nature des responsabilités qui ont pu conduire à ce genre d'évènement"*, écrit-il dans un communiqué. *"Certes, nous devons encore apprendre à vivre avec les algorithmes, mais l'usage d'une technologie, quelle qu'elle soit, ne peut en rien permettre aux éditeurs de contenus de se soustraire à leur propre responsabilité."* L'Union européenne travaille depuis deux ans sur l'IA Act, un texte visant à encadrer l'utilisation de l'intelligence artificielle. Mais le secrétaire d'État veut se saisir du texte pour renforcer *"le niveau de risque de certaines applications de l'IA"*, pour mieux sensibiliser et protéger les utilisateurs. Pour ce faire, M. Michel annonce avoir mis en place un groupe de travail *"afin d'analyser le texte en cours de préparation auprès de l'UE et de proposer des adaptations indispensables"*. **P.-F.L.**