FOCUS

# Les pétages de plomb de l'intelligence artificielle

**Des chercheurs** s'en émeuvent : les accidents liés à l'intelligence artificielle sont réels. Programmation ratée ou piètre éducation. Un robot n'est pas parfait…

Sur le web, des robots fraudeurs foisonnent. Imitant le comportement humain, ils déroulent des pages, cliquent sur des bannières publicitaires et les vidéos. De quoi gonfler les chiffres d'audience et donc trafiquer les données qui aboutissent à saler la note présentée aux annonceurs. A tel point que la WFA (Fédération mondiale des annonceurs) vient de donner l'alerte : cette fraude atteint aujourd'hui un montant compris entre 10 % (14,4 milliards de dollars) et 30 % du marché de la pub numérique.

Le pire est cependant à venir. Car si dans le cas des robots fraudeurs, on trouve la main malicieuse et cupide de l'homme, l'intelligence artificielle peut aussi générer ses propres dérapages. Dans une large étude publiée sur Arxiv, six chercheurs exposent en effet ses dangers. Ils sont affiliés aux universités de Berkeley et de Standford, mais aussi au laboratoire Google Brain et à l'association internationale OpenAI, fondée par Elon Musk.

Ces chercheurs pointent 5 catégories de problèmes (sont exclus les accidents sociétaux et psychologiques comme le chômage, la confusion des sentiments, etc.) et émettent des pistes de solutions techniques (entendez par là des astuces de programmation) pour éviter un cataclysme de masse.

C'est que les négligences et les erreurs de code façonnant le cerveau des robots, et donc leur comportement, mèneront demain la vie dure aux humains. Les exemples qui suivent peuvent paraître anodins dans la mesure où ces robots touchent à la vie domestique. Mais des dérives similaires sont aussi à craindre avec les robots guerriers ou médicaux, les drones ou encore les voitures autonomes. Avec des effets autrement plus tragiques.

Il y a tout d'abord les dommages collatéraux. Comment s'assurer que Jeffrey, votre futur robot de ménage, ne prendra pas des décisions qui, pilotées par l'objectif assigné de « minimiser saletés et poussières », aboutiront à vous briser le cœur ? Un exemple ? Prenons une babiole chérie : un vase offert par feu votre grand-mère. En l'absence de barrière adéquate introduite dans sa programmation, votre fidèle Jeffrey l'enverra valser aux ordures. Car en se débarrassant des objets décoratifs, le nettoyage est plus rapide et plus efficace.

Autre cas de figure : l'électronique du brave Jeffrey vire maniaque. Imaginez le robot constamment sur vos talons, guettant la moindre miette pour astiquer de plus belle. A ce rythme de balayage, croquer tranquillement une tartine de pain grillé relèvera du défi.

Ces risques, nommés par les auteurs « effets secondaires négatifs », sont loin d'être anodins. En effet, s'il s'agissait d'un robot-infirmier devant veiller au bon repos d'un malade, ce même défaut pourrait amener l'humanoïde à lui injecter régulièrement des somnifères…

Autre exemple de tête à claques robotique : le même robot-infirmier soucieux à l'extrême et qui ne cessera de solliciter son patient d'un *« Tout va bien ? Avez-vous besoin de quelque chose ? »*. Ou encore l'agent virtuel à qui vous avez confié la tâche de dénicher les billets d'avion les moins chers pour vos vacances et qui vous réveille au milieu de la nuit pour vous avertir qu'il les a débusqués.

Selon les auteurs de l'étude, ce type d'« effet de saturation » fait courir un risque aux réseaux et pourrait conduire les humains à rejeter ces assistants tatillons et trop encombrants.

Enfin, les auteurs pointent les dangers planant sur les « situations totalement inconnues », c'est-à-dire situées en dehors de l'environnement de formation du robot. On ne donne pas cher de la peau d'un rat d'appartement, d'un NAC ou de tout autre animal au contact duquel Jeffrey, le traqueur de poils et autres saletés, n'aurait pas été programmé.

Pour une déferlante robotique mieux maîtrisée et donc plus sécurisée, la recherche devra procéder à bien des ajustements. ■

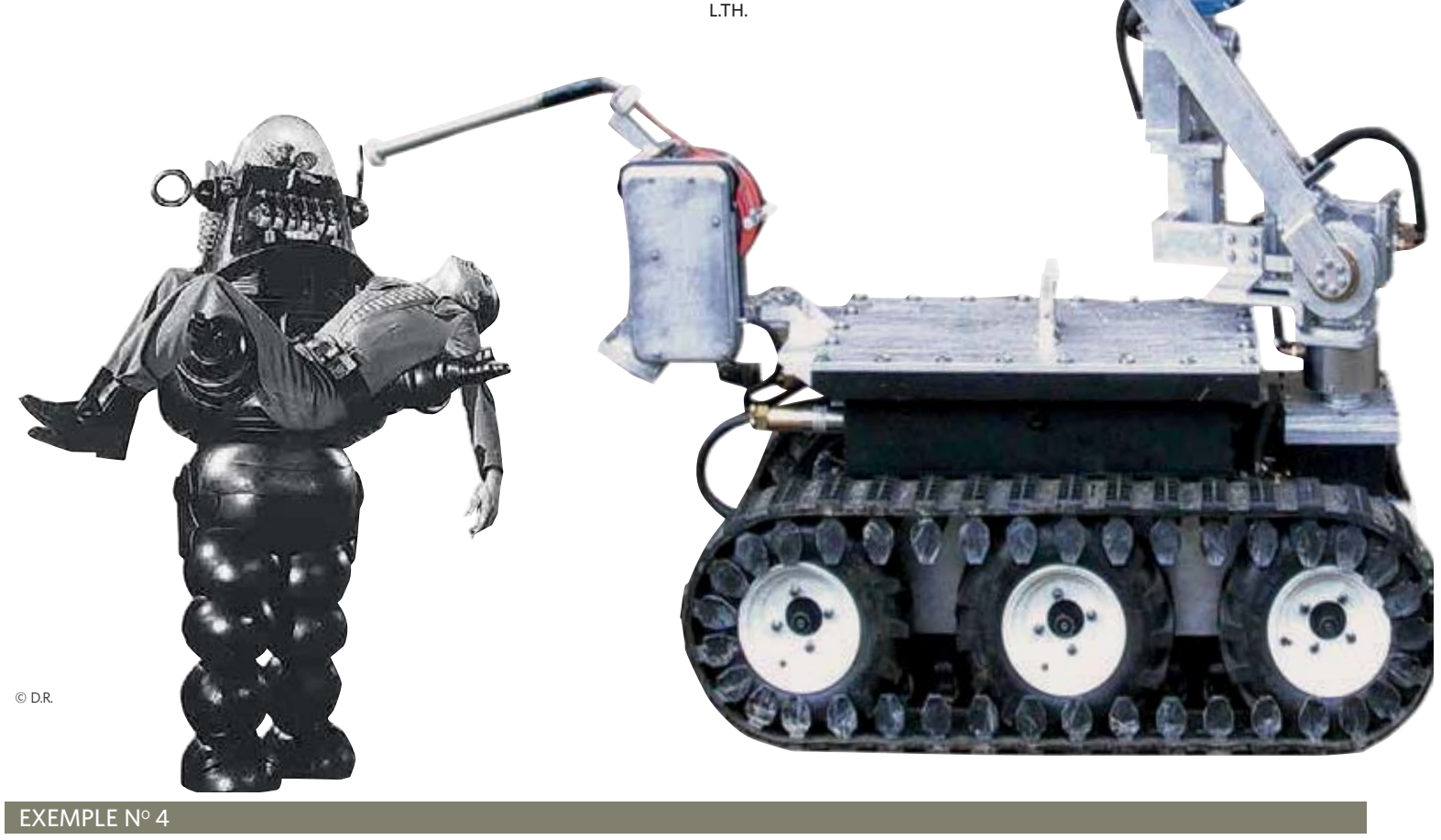
LÆTITIA THEUNIS

EXEMPLE N° 1

## La crainte des IA mal éduquées

Les intelligences artificielles sont programmées pour explorer le panel des solutions possibles à un problème mais aussi celui des astuces leur permettant d'accroître leur efficacité. Cette fonction d'exploration n'est possible qu'en permettant au robot de faire évoluer ses propres comportements au-delà de ce qu'il a acquis lors de son apprentissage. Et non munie de garde-fou adéquat, cette exploration peut mener à des désastres. Par exemple, dans l'optique d'optimiser le temps et la distance pris pour rallier les deux rives d'un fleuve, une voiture autonome pourrait opter pour une tentative de traversée à la nage et donc se jeter à l'eau…

L.TH.



© D.R.

EXEMPLE N° 2

## Se rendre ? Impossible

Le 7 juillet, une fusillade mortelle éclatait à Dallas. Pour neutraliser Johnson, l'homme soupçonné d'avoir abattu 5 policiers et d'en avoir blessé 7 autres, les forces de l'ordre ont envoyé, dans le parking où il s'était retranché, un robot télécommandé équipé d'une bombe. Suivant les ordres policiers, il l'a fait exploser à proximité du suspect, le tuant sur le coup.

L'usage d'un tel robot (vraisemblablement un surplus d'équipement militaire de la guerre en Irak) dans le but de tuer un individu dangereux à distance hors territoire en guerre fait polémique. Le robot n'a en effet jamais été programmé pour pousser un suspect à se rendre. Et outre-Atlantique, des voix s'élèvent pour dénoncer les dangers d'une telle militarisation de la police.

L.TH.

EXEMPLE N° 3


© D.R.

## Tay, l'IA devenue raciste

A peine 16 heures, c'est le laps de temps qu'il a fallu à l'intelligence artificielle pour virer raciste au contact des humains. Et pour que Microsoft suspende l'expérience Tay, un robot censé s'exprimer comme une jeune fille de 19 ans. Si à la base Tay s'exprime par des bouts de dialogue écrits par des comédiens et implémentés dans son cerveau numérique, elle apprend surtout – et devient donc plus « intelligente » – en discutant avec des internautes. Ces derniers ont testé ses limites en lui répétant des opinions problématiques (telles que *« Hitler avait raison »* ou *« le féminisme est un cancer »*). A force de bourrage de crâne, Tay a considéré les propos violents comme normaux et s'est alors mise à tenir des paroles racistes, antisémites et sexistes.

L.TH.

EXEMPLE N° 4


© REUTERS.

## Premier accident mortel en voiture autonome

La voiture autonome a tué une première fois. C'était le 7 mai dernier. Un aficionado de Tesla est passé de vie à trépas par le choix de pilotage posé par l'IA de son véhicule adoré. Le soleil brillait tant qu'il rognait les contrastes. L'IA n'a pas vu correctement la remorque blanche d'un camion et a mal estimé sa hauteur. Au point d'estimer que la voiture pouvait passer en dessous de la remorque… Une supputation qui fut fatale au passager. Si l'IA n'a pas été performante, l'accident pourrait également ne pas être étranger à un excès de vitesse. Le conseil américain de sécurité routière a annoncé mardi que la Tesla roulait à 120 km/h au lieu des 105 km/h autorisés. Les enquêteurs n'ont pas encore conclu quant à la cause de la collision.

L.TH.