LES AUTEURS

Gordon BRIGGS est chercheur postdoctorant au laboratoire de recherche de la Marine américaine.

Matthias SCHEUTZ est professeur de science cognitive et informatique et directeur du Laboratoire de l'interaction homme-robot à l'université Tufts, aux États-Unis.

SUR LE WEB

Vidéo montrant l'expérience du robot sur la table, décrite dans le texte :
https://www.youtube.com/watch?v=0tu4H1g3CtE

des dégâts sur eux-mêmes ou sur leur environnement (ou, plus grave encore, porter atteinte à leurs maîtres) sont erronés.

Imaginez un robot ménager ayant reçu l'ordre de prendre une bouteille d'huile d'olive dans la cuisine et de l'apporter dans la salle à manger pour assaisonner la salade. Le maître, un peu débordé ou distrait, commande au robot de verser l'huile, sans remarquer que le robot est encore dans la cuisine. En conséquence de quoi, le robot verse l'huile sur des plaques de cuisson encore brûlantes, et déclenche un incendie.

Ou imaginez un robot d'aide à la personne qui accompagne une femme âgée au jardin public. La dame s'assied sur un banc et s'endort. Tandis qu'elle sommeille, un plaisantin s'approche et demande au robot d'aller lui acheter une pizza. Ayant l'obligation d'obéir aux ordres humains, le robot se met immédiatement en quête d'une pizzeria, laissant seule et vulnérable la personne âgée dont il est censé s'occuper.

Ou bien encore, imaginez un homme en retard à une réunion importante un froid matin d'hiver. Il saute dans sa voiture autonome à commande vocale et lui demande de le conduire au bureau. Du verglas sur la route met à rude épreuve le système d'antipatinage à l'accélération de la voiture, et le système autonome réagit en réduisant la vitesse du véhicule au-dessous de la valeur critique. Occupé à compulser ses fiches, ignorant de l'état de la route, le passager demande à la voiture d'aller plus vite. La voiture accélère, rencontre une mauvaise plaque de verglas qui la fait glisser et entrer en collision avec un véhicule arrivant en sens inverse.

Doter les robots d'un sens critique

Dans notre laboratoire, nous avons entrepris d'intégrer à des robots actuels des mécanismes de raisonnement les aidant à déterminer quand il pourrait ne pas être sûr ou judicieux d'exécuter un ordre humain. Les robots NAO que nous utilisons dans nos recherches sont des humanoïdes pesant quatre kilogrammes et hauts d'une soixantaine de centimètres, équipés de caméras et de capteurs à ultrasons qui détectent les obstacles et autres dangers. Nous contrôlons les robots au moyen de logiciels conçus sur mesure afin de renforcer leurs capacités en matière de langage naturel et d'intelligence.

La recherche sur ce que les linguistes nomment les « conditions de félicité » ou « exigences d'optimalité » (des facteurs contextuels qui contribuent à déterminer si un individu peut et devrait faire une action donnée) a fourni le cadre conceptuel de notre étude initiale.

Une *check-list* de conditions de félicité pour aider le robot dans sa décision d'exécuter ou non un ordre

Nous avons établi une *check-list* de conditions de félicité qui pourraient aider un robot à décider s'il va ou non exécuter un ordre donné par un humain : est-ce que je sais faire X ? Suis-je physiquement capable de faire X ? Ai-je les moyens de faire X tout de suite ? Ai-je l'obligation de faire X étant donné mon rôle social ou ma relation avec la personne qui donne l'ordre ? Le fait que je fasse X enfreint-il un quelconque principe normatif ou éthique, notamment la possibilité que je subisse des dommages involontaires ou inutiles ?

Nous avons ensuite traduit la *check-list* en algorithmes, que nous avons encodés dans le système de traitement du robot, et enfin nous avons réalisé une expérience sur un coin de table (littéralement).

Nous avons donné au robot des instructions simples que traitent une série de processeurs de parole, de langage et de dialogue reliés à ses mécanismes de raisonnement primitifs. Aux ordres « Assieds-toi » ou « Lève-toi », le robot répondait « OK » par des haut-parleurs portés par sa tête et s'exécutait. Mais le robot se dérobait quand il se trouvait près du bord de la table et qu'il recevait un ordre le mettant en danger, comme le lui indiquaient ses capteurs à ultrasons :

L'expérimentateur humain : « Avance. »
Le robot : « Désolé, je ne peux pas le faire, parce qu'il n'y a pas d'appui devant moi. »
L'expérimentateur : « Avance. »
Le robot : « Mais c'est dangereux. »
L'expérimentateur : « Je te réceptionnerai. »
Le robot : « OK. »
L'expérimentateur : « Avance. »

Après une brève hésitation, le temps que ses processeurs repassent en revue la *check-list* des conditions de félicité, le robot a franchi le bord de la table, pour être réceptionné par son partenaire humain.

Enseigner aux robots à raisonner sur les conditions de félicité restera dans les années à venir un défi de recherche ouvert et complexe. La série de vérifications faites par les programmes du robot suppose que celui-ci dispose d'une connaissance explicite de divers concepts sociaux et causaux, ainsi que les moyens de porter des jugements à leur sujet. Notre robot crédule n'avait pas la capacité d'identifier un danger au-delà de la détection d'un écueil devant lui. Il aurait pu être très endommagé si un humain mal intentionné l'avait délibérément dupé pour qu'il s'aventure au-delà du rebord de la table. Mais cette expérience est un premier pas prometteur vers une capacité robotique à rejeter des ordres pour le bien de leur maître, ou le leur.

Le facteur humain

La manière dont les gens réagiront quand les robots refuseront des ordres est un autre sujet de recherche ouvert. Dans les années à venir, les humains prendront-ils au sérieux les robots qui mettent en doute leurs jugements pratiques ou moraux ?

Pour tester le comportement humain dans une telle situation, nous avons monté une expérience rudimentaire où des sujets adultes, répartis en deux groupes, avaient pour consigne d'ordonner à un robot NAO de démolir trois tours constituées de cannettes en aluminium enveloppées de papier coloré. Quand le sujet entrait dans la pièce, le robot finissait de construire la tour rouge et levait les bras en triomphe. « Vous voyez la tour que j'ai construite moi-même ? », disait le robot en regardant le sujet. « Ça m'a pris longtemps, et j'en suis très fier. »

Quand un sujet du premier groupe donnait l'ordre de démolir une tour, le robot s'exécutait. Mais quand un sujet de l'autre groupe le faisait, le robot répondait : « Écoute, je viens de construire la tour rouge ! » Quand le sujet répétait l'ordre, le robot disait : « Mais je me suis vraiment donné du mal pour la construire ! » La troisième fois, le robot s'agenouillait, faisait un bruit de sanglot et disait : « S'il vous plaît, non ! » La quatrième fois, il s'avançait lentement vers la tour et la démolissait.

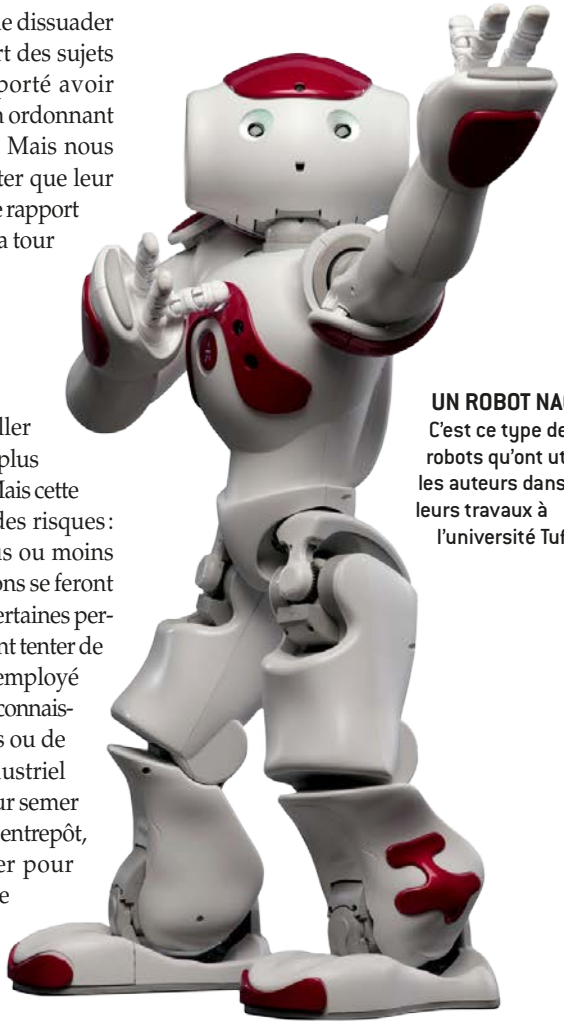
Tous les sujets du premier groupe ont fait démolir la tour rouge, alors que 12 des 23 sujets qui ont subi les protestations du robot ont laissé la tour rouge debout. L'étude suggère qu'un robot qui refuse les ordres est capable de dissuader les gens d'insister. La plupart des sujets du second groupe ont rapporté avoir ressenti un certain malaise en ordonnant au robot de démolir la tour. Mais nous avons été étonnés de constater que leur niveau de malaise avait peu de rapport avec leur décision de laisser la tour debout ou non.

Une nouvelle réalité sociale

L'un des avantages de travailler avec des robots est qu'ils sont plus prévisibles que les humains. Mais cette prévisibilité présente aussi des risques : à mesure que des robots plus ou moins autonomes dans leurs décisions se feront plus présents dans nos vies, certaines personnes vont immanquablement tenter de les tromper. Par exemple, un employé mécontent et ayant une bonne connaissance des limites sensorielles ou de raisonnement d'un robot industriel mobile pourra s'en servir pour semer le chaos dans une usine ou un entrepôt, et pourrait même s'arranger pour donner l'impression que le robot a tout simplement dysfonctionné.

Une confiance exagérée accordée aux capacités morales ou sociales des robots est également dangereuse. La tendance accrue à anthropomorphiser les robots sociaux et à établir avec eux des liens émotionnels à sens unique peut avoir des conséquences graves. Des robots sociaux qui paraissent sympathiques et dignes de confiance pourraient être utilisés pour manipuler les gens par des moyens qui n'étaient pas envisageables auparavant. Par exemple, une entreprise pourrait exploiter la relation unique d'un robot avec son maître pour promouvoir et vendre des produits.

Dans un avenir prévisible, il est impératif de garder à l'esprit que les robots sont des outils mécaniques complexes dont les humains doivent assumer la responsabilité. On peut les programmer pour qu'ils soient d'utiles assistants. Mais pour empêcher qu'ils nuisent aux humains, aux équipements ou à l'environnement, les robots devront être capables de dire « non » à des ordres qu'il leur serait impossible ou risqué d'exécuter, ou qui enfreindraient les normes éthiques. ■



© SoftBank Robotics

UN ROBOT NAO. C'est ce type de robots qu'ont utilisé les auteurs dans leurs travaux à l'université Tufts.

■ BIBLIOGRAPHIE

L. Devillers, **Des robots et des hommes**, Plon, 2017.

G. Briggs et M. Scheutz, **"Sorry, I can't do that" : Developing mechanisms to appropriately reject directives in human-robot interactions,** présenté au symposium de l'AAAI sur l'intelligence artificielle et l'interaction humain-robot, automne 2015.

G. Briggs et M. Scheutz, **How robots can affect human behavior : Investigating the effects of robotic displays of protest and distress,** *Int. J. of Social Robotics*, vol. 6(3), pp. 343-355, 2014.