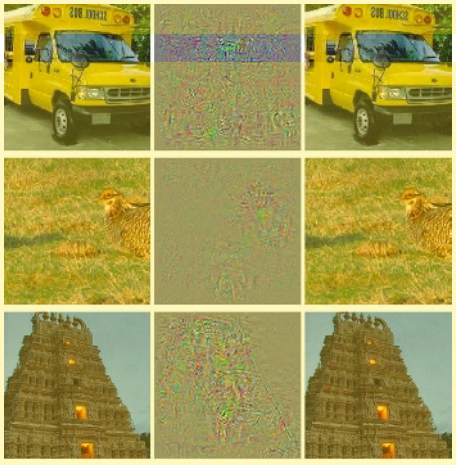
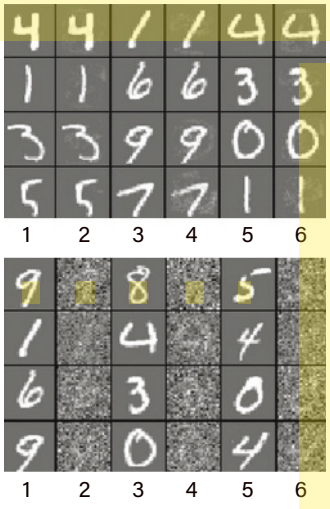
DES IMAGES PIÈGES

Un procédé de modification d'images permet de piéger les algorithmes neuronaux de reconnaissance d'images et les conduit à produire des réponses absurdes.



Les images proposées dans la colonne de gauche de la figure *ci-contre*, après la phase d'apprentissage du réseau de neurones, sont toutes correctement reconnues. Christian Szegedy et son équipe calculent alors des perturbations *(colonne centrale)* qui, appliquées aux images de la colonne de gauche, donnent les images de la colonne de droite. Ces dernières sont alors, dans chaque cas, reconnues par le système comme figurant… des autruches !



Dans le premier tableau de la figure *ci-contre,* les chiffres manuscrits des colonnes 1, 3, et 5 sont identifiés correctement par le réseau de neurones (après apprentissage). En revanche, le réseau identifie mal les mêmes images qui ont subi de légères modifications (colonnes 2, 4 et 6), alors que l'œil humain les identifie aussi bien et ne voit presque pas de différences avec les premières images. Dans le second tableau, on a placé en colonnes 2, 4 et 6 les images des colonnes 1, 3 et 5 soumises à un assez fort bruit aléatoire. Le réseau continue pourtant de fournir les bonnes identifications dans plus de la moitié des cas, alors que les humains en sont incapables.

Ces deux tableaux montrent donc que ce qu'un réseau de neurones apprend dans sa phase d'apprentissage est d'une nature assez différente de ce que nous-mêmes apprenons.

Les chercheurs du domaine étaient persuadés qu'un réseau ayant bien appris était nécessairement robuste et doté d'une propriété de continuité : des traits dessinant de manière vague et approchée une lettre seront reconnus correctement, et une image proche d'une image reconnue sera elle aussi reconnue.

Or ce n'est pas le cas : les chercheurs ont mis au point des images pièges, conçues pour qu'un réseau ayant appris à reconnaître une certaine catégorie d'images propose une réponse incongrue et fausse quand on lui soumet une image proche, qu'un humain classe aisément *(voir l'encadré ci-contre)*.

Insistons sur la prudence nécessaire. Yann LeCun, pionnier de l'« apprentissage profond », qui dirige la recherche en intelligence artificielle chez Facebook, est clair : « C'est un abus de langage que de parler de neurones ! De la même façon qu'on parle d'aile pour un avion comme pour un oiseau, le neurone artificiel est un modèle extrêmement simplifié de la réalité biologique. » Il ajoute : « Les intelligences artificielles les plus abouties ont aujourd'hui moins de sens commun qu'un rat ! »

En somme, les réussites des réseaux de neurones, des méthodes d'apprentissage automatique et des méthodes de fouille de données massives (les *big data*) sont de nouveaux exploits des machines, mais ces succès ne signifient pas que les machines sont devenues véritablement intelligentes.

Revenons à nos exemples pièges. Une équipe de sept chercheurs affiliés aux universités de New York et Montréal, à Google et à Facebook, réunie autour de Christian Szegedy, a publié en 2014 une méthode permettant de tromper les systèmes de reconnaissance visuelle.