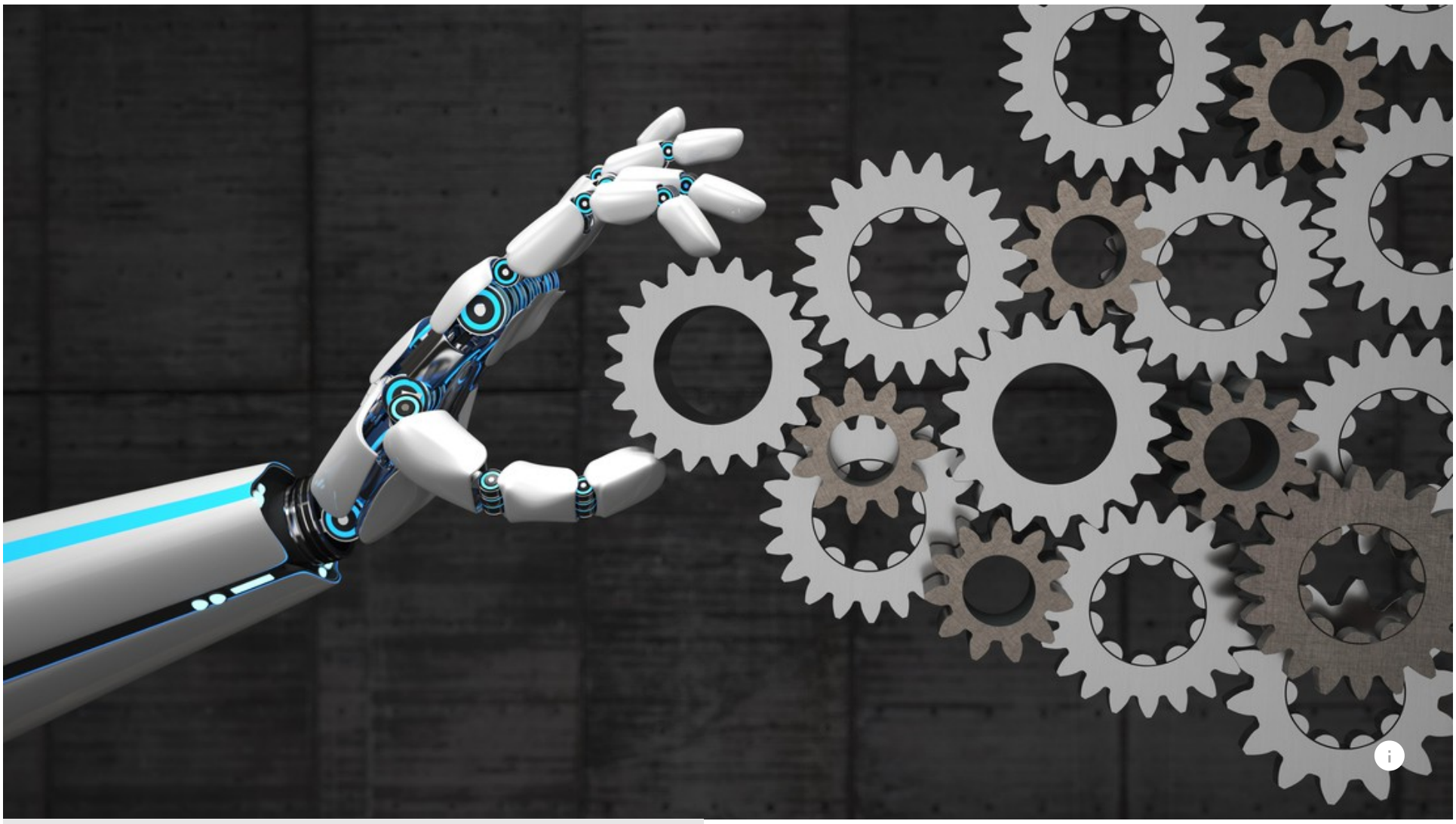
L'éthique et l'IA: un principe d'optimisation non éthique



Le professeur Anthony Davison de l'EPFL et ses co-auteurs présentent une base mathématique pour les préoccupations qui entourent les implications éthiques de l'IA.

L'intelligence artificielle (IA) est de plus en plus couramment employée tout autour de nous et pourrait offrir d'importants avantages potentiels. Mais l'utilisation non éthique de l'AI fait aussi naître des préoccupations croissantes. Le professeur Anthony Davison, qui occupe la Chaire de statistique à l'EPFL, et des collègues britanniques se sont penchés sur ces questions d'un point de vue mathématique en se focalisant sur l'IA commerciale ayant pour objet de maximiser les profits.

01.07.20

MOTS-CLÉS

Anthony C. Davison
informatique
intelligence artificielle
mathématiques
SB
Sciences de Base

ACTUALITÉS

- Toutes les actualités EPFL
- Toutes les actualités

ABONNEMENT

Recevez un e-mail à chaque publication

Songeons au cas d'une société d'assurance recourant à l'IA dans le but de trouver une stratégie pour décider des primes à appliquer aux clients potentiels. L'IA choisira parmi des stratégies potentielles et certaines d'entre elles pourraient être discriminatoires ou faire un usage impropre des données clients susceptible d'entraîner ultérieurement de lourdes sanctions financières pour l'entreprise. Idéalement, des stratégies non éthiques telles que celles-ci devraient être écartées par anticipation du paysage de stratégies potentielles. Hélas, l'IA est dénuée de sens moral et ne peut dès lors distinguer les stratégies éthiques de celles qui ne le sont pas.

Dans un article publié dans *Royal Society Open Science* en date du 1er juillet 2020, Davison et ses co-auteurs Heather Battey (Imperial College de Londres), Nicholas Beale (Sciteb Limited) et Robert MacKay (Université de Warwick) montrent que l'IA pourrait opter pour une stratégie non éthique dans un grand nombre de situations. Ils formulent leurs résultats sous la forme d'un «Principe d'optimisation non éthique»:

Si l'IA a pour but de maximiser un rendement ajusté au risque, dans des conditions normales, elle risque de manière disproportionnée d'opter pour une stratégie non éthique, sauf si la fonction objective tient suffisamment compte de ce risque.

Ce principe peut aider les gestionnaires de risque et les régulateurs notamment à déceler les stratégies non éthiques que pourrait recéler le vaste paysage de stratégies potentielles. Idéalement, on devrait pouvoir configurer l'IA de manière à éviter les stratégies non éthiques, mais cela pourrait se révéler impossible, parce qu'il est impossible de les définir à l'avance. Pour pouvoir piloter l'utilisation de l'IA, l'article indique comment estimer la proportion de stratégies non éthiques et la distribution des stratégies les plus efficaces.

«Nos travaux pourraient aider les régulateurs et les personnes chargées du respect de la compliance notamment à identifier les stratégies problématiques que pourrait recéler le vaste paysage de stratégies potentielles. Ce paysage devrait contenir un nombre disproportionnellement élevé de stratégies non éthiques. L'analyse de ce paysage devrait indiquer où des problèmes risquent de surgir et suggérer dès lors en quoi l'algorithme de recherche de l'IA devrait être modifié pour les éviter», déclare le professeur Davison. «Cela tend aussi à démontrer qu'il pourrait être nécessaire de repenser le mode de fonctionnement de l'IA dans de très grands paysages de stratégies, de façon à ce que les résultats non éthiques soient explicitement rejetés durant le processus d'apprentissage.»

Partager sur

Madame Wendy Hall, directrice de l'Ada Lovelace Institute au Royaume-Uni, qui cherche à garantir que les données et l'IA soient une force positive dans la société, a déclaré pour sa part: «C'est un article vraiment important. Il montre que nous ne pouvons pas attendre des systèmes d'IA qu'ils agissent de façon éthique parce que leurs objectifs semblent éthiquement neutres. Au contraire, dans des conditions normales, un système d'IA optera de façon disproportionnée pour des solutions non éthiques, sauf s'il a été soigneusement conçu pour les éviter.

Les formidables avantages potentiels de l'IA ne pourront être correctement exploités que si le comportement éthique a été pris en compte dès le départ dans le processus de conception, en intégrant ce «Principe d'optimisation non éthique» dans une diversité de perspectives. Fort heureusement, ce principe peut aussi être utilisé pour déceler des problèmes éthiques dans des systèmes existants et les résoudre ensuite en améliorant la conception.»

Financement
Fonds national suisse de la recherche scientifique (FNS), UK Engineering and Physical Science Research Council, Alan Turing Institute, Capital International

Références
Beale N, Battey H, Davison AC, MacKay RS. (2020) An unethical optimization principle. R. Soc. Open Sci. 7: 200462.
http://dx.doi.org/10.1098/rsos.200462

Auteur: Orane Jecker

Source: EPFL

Connexion

EPFL

Contact EPFL CH-1015 Lausanne +41 21 693 11 11

Suivre les pulsations de l'EPFL sur les réseaux sociaux

Accessibilité Mentions légales

© 2020 EPFL, tous droits réservés