Human compatible

Type de contenu : Texte

Type de médiation : sans médiation

Type de support : Volume

Titre(s) : Human compatible : artificial intelligence and the problem of control / Stuart Russell

Auteur(s) : Russell, Stuart (1962-....)

Publication : [New York (N.Y.)] : Penguin books, 2020

Description matérielle : 1 vol. (XIV-336 p.) : ill. ; 22 cm

ISBN : 0-525-55863-2
978-0-525-55863-7

EAN : 9780525558637 br.

Classification décimale Dewey : 006.301

Note sur les bibliographies et les index : Bibliogr. pp.([299]-323). Index

Résumé ou extrait : Présentation de l'éditeur : "In the popular imagination, superhuman artificial intelligence is an approaching tidal wave that threatens not just jobs and human relationships, but civilization itself. Conflict between humans and machines is seen as inevitable and its outcome all too predictable. In this groundbreaking book, distinguished AI researcher Stuart Russell argues that this scenario can be avoided, but only if we rethink AI from the ground up. Russell begins by exploring the idea of intelligence in humans and in machines. He describes the near-term benefits we can expect, from intelligent personal assistants to vastly accelerated scientific research, and outlines the AI breakthroughs that still have to happen before we reach superhuman AI. He also spells out the ways humans are already finding to misuse AI, from lethal autonomous weapons to viral sabotage. If the predicted breakthroughs occur and superhuman AI emerges, we will have created entities far more powerful than ourselves. How can we ensure they never, ever, have power over us? Russell suggests that we can rebuild AI on a new foundation, according to which machines are designed to be inherently uncertain about the human preferences they are required to satisfy. Such machines would be humble, altruistic, and committed to pursue our objectives, not theirs. This new foundation would allow us to create machines that are provably deferential and provably beneficial. In a 2014 editorial co-authored with Stephen Hawking, Russell wrote, "Success in creating AI would be the biggest event in human history. Unfortunately, it might also be the last." Solving the problem of control over AI is not just possible; it is the key that unlocks a future of unlimited promise" ; In the popular imagination, conflict between humans and machines is seen as inevitable and its outcome all too predictable. Russell argues that this scenario can be

avoided, but only if we rethink AI from the ground up. He explores the idea of intelligence in humans and in machines, describes the near-term benefits we can expect from intelligent personal assistants and accelerated scientific research, and outlines the AI breakthroughs that still have to happen before we reach superhuman AI. Russell also spells out the ways humans are already misusing AI, from lethal autonomous weapons to viral sabotage."

Sujet - Nom commun : Intelligence artificielle -- Société
Automatisation -- Mesures de sûreté