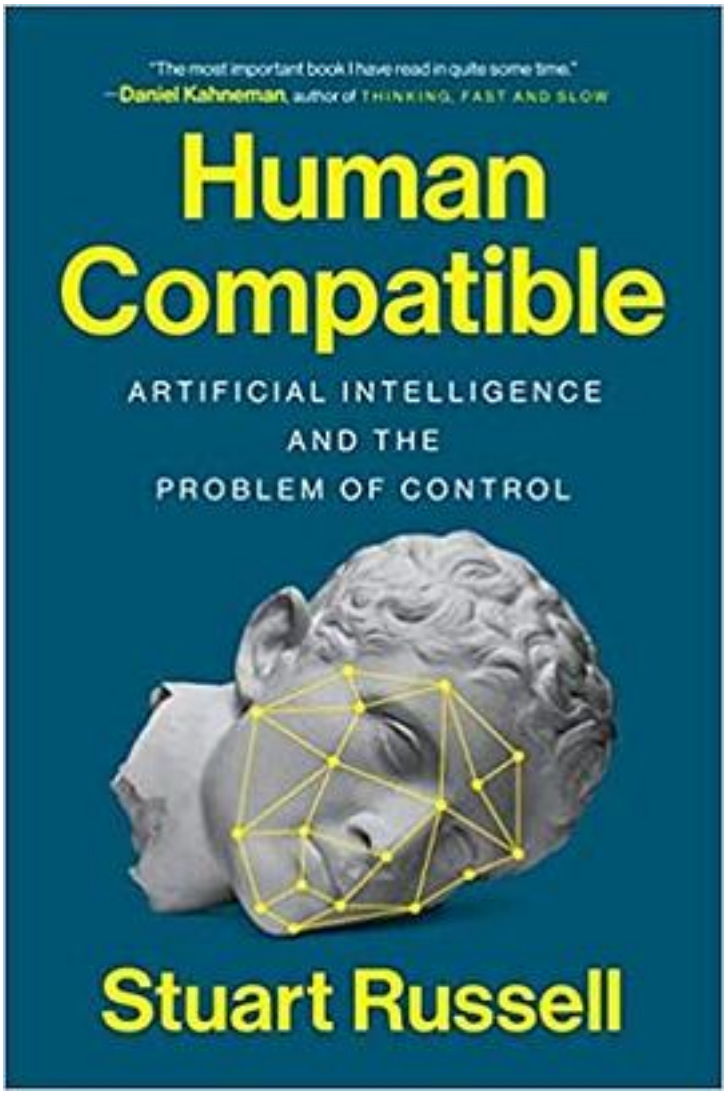
Human Compatible



Human Compatible_下载链接1_

著者:Stuart Russell

出版者:Viking

出版时间:2019-10-8

装帧:Hardcover

isbn:9780525558613

A leading artificial intelligence researcher lays out a new approach to AI that will enable us to coexist successfully with increasingly intelligent machines

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.

作者介绍:

Stuart Russell is a professor of Computer Science and holder of the Smith-Zadeh Chair in Engineering at the University of California, Berkeley. He has served as the Vice-Chair of the World Economic Forum's Council on AI and Robotics and as an advisor to the United Nations on arms control. He is a Fellow of the American Association for Artificial Intelligence, the Association for Computing Machinery, and the American Association for the Advancement of Science. He is the author (with Peter Norvig) of the definitive and universally acclaimed textbook on AI, Artificial Intelligence: A Modern Approach.

目录:

Human Compatible_下载链接1_

标签

人工智能

机器学习

价值观

AI

评论

此书作者是AI领域的大牛，那本至今为止可谓影响最大的教科书便是他的杰作，然而我从原本打算给两星评价上调到三星就只是因为其名声太大，敬畏之。说到书的质量，呵呵，这是写给思想家、哲学家、历史学家、资本运作方甚至未来学家们读的书，它太过于高大上，给我一种何不食肉糜的感觉。我个人对一切脱离实际代码、算法与数据的议论都有偏见并避而远之，你不要和我扯什么历史发展的轨迹、伟人过往的言论、个案堆砌出来的推断，我要的是实实在在的代码演示、算法潜力分析与数据实操技术，能让我在近期就着手推动改进项目，而这就是我更推崇Marcus新书的原因，虽然那本书也没有对AI的各种弱点提出具体解决方案，可其回顾的东西具备一定程度的可操作性，而这书却单纯像兴奋剂，high过之后就只剩空虚。非AI专业完全不推荐此书，除非用来励志。

Stuart Russell教授的最新著作，讨论这个时代最重要的问题之一：面对可能比我们更聪明的机器时，人类的未来命运。一个有益的超级人工智能ASI如果真能实现，将会给世界带来什么？虽然可以做很多科幻猜想，但一个较低的下限是，人工劳动可能会消失，一切应该都非常便宜。当世界的GDP每年都增长十倍，也许相互争夺不再有意义。但这些都假定我们能够控制ASI，我们应该对此保持谨慎。机器智能的标准模型中，通常假定一些明确已知的目标，但如果我们设定了错误的目标，那么机器将无情地追求，并导致我们不希望的结果。因此，本书中提出了AI系统设计的三原则：1）机器的唯一目标是最大程度地实现人类的偏好；2）机器最初并不清楚这些偏好是什么；3）关于人类偏好的最终信息源是人类的行为。并提出逆向强化学习，试图寻找解决方案。

纽约回国飞机上终于读完了 拖拖拉拉一个月了

Human Compatible_下载链接1_

书评

Human Compatible_下载链接1_