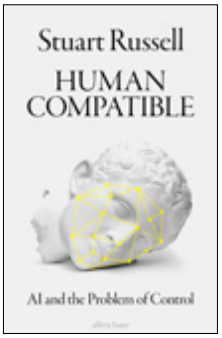
Book Review

Human Compatible: AI and the Problem of Control



Author: Stuart Russell

Publisher: Allen Lane; 1st edition (8 Oct. 2019) (**ISBN:** 978-0241335208), 352 pages

Reviewed by Group Captain Emma Keith

Introduction

Human Compatible considers the interface between humans and machines and asks the question of how we ensure that the AI we are creating is compatible with the world that humans want to live in. The author, Stuart Russell, is a professor of Computer Science at UC Berkeley, a published author on this topic and a recognised leading expert in artificial intelligence and machine learning. He is not an outside alarmist, he is at the very heart of AI and machine research and presents his concerns in an intelligent, persuasive, and logical way. The book is accessible to a general audience and provides a brilliant introduction to some of the more strategic issues associated with the rapid development and global investment in AI. It will equally appeal to the AI specialist to challenge them to view AI in alternate ways.

In clear and compelling language, Stuart Russell provides a unique perspective on the interaction between humans and machines, he draws on the past, present and the future to present a holistic and balanced view. The book is divided into three main sections, the first, explores ideas of intelligence in humans and in machines and it unpicks the definition of intelligence by questioning what we mean by intelligence? One comment that stood out was, 'Humans are intelligent to the extent that our actions can be expected to achieve our objectives. Machines are intelligent to the extent that their actions can be expected to achieve their objectives (p. 9).' Stuart Russell explores the concept that the vital component is in defining what the objectives are – for example, giving a machine the objective of saving the planet, might mean that it decides to delete the biggest risk to the planet – i.e. the human race.

Clearly highly knowledgeable in this area the author presents the content in easily accessible language and there is no technical knowledge required. I did, however, find the added appendix beneficial for when I wanted to understand a concept in greater depth.

The second section is a fascinating look at some of the problems arising from imbuing machines with intelligence with a particular focus on the issue of control: retaining absolute power over machines that are more powerful than us. Comments such as, 'we have to face the fact that we are planning to make entities that are far more powerful than humans. How do we ensure that they never, ever have power over us? (p. 8)', certainly got my attention. This whole section really captured my imagination as the content cleverly weaved between technical detail and philosophical approaches to life and reinforced how we must consider if what we currently wish for from machines will turn out to be what we really want? It was fascinating to read the authors thoughts on how an intelligent machine may ensure it cannot be turned off as part of its desire to achieve its objective. This was not something I had previously considered, always assuming that an 'off' switch was a way out. The book also reflects on the amount of damage currently caused by basic algorithms nudging human behaviour via social media, for example, and ponders what truly highly intelligent AI could do?

The book concludes with a final section that suggests a new way to think about AI to ensure that machines remain beneficial to humans, forever. Stuart Russell summarises, 'Everything civilization has to offer is the product of our intelligence; gaining access to considerably greater intelligence would be the biggest event in human history. The purpose of the book is to explain why it might be the last event in human history and how to make sure that it is not (Preface).'

I found this a captivating read and regularly read sections out to those around me as the points made were profound and thought provoking. It really made me stop and think. The book does a brilliant job of making a technical and serious topic accessible, easy to engage with, and genuinely gripping. I also enjoyed his dry sense of humour and witty side comments that had the ability to make you laugh whilst simultaneously cleverly reinforcing a point. I highly recommend it to anyone who wishes to engage and understand the world in which we live in, arguably, it is essential reading for anyone who cares about our future.