****HUMAN COMPATIBLE**

**AI and the Problem of Control**

**By Stuart Russell**

**8 October 2019 I Allen Lane I £25.00**

**\*Stuart Russell will be in the UK Mon 7- Fri 11 October 2019\***

**The world’s leading AI expert explains why he has come to consider his own discipline an existential threat to his own species, and lays out how we can change course before it’s too late.**

In *Human Compatible* Russell explains why uncontrolled superhuman AI is the most important issue facing mankind in this century. He lays out a radically new direction to designing AI that will enable humanity to coexist successfully with increasingly competent machines. Above all, they must allow themselves to be switched off (the Hal problem). If their learning behaviours can be made compatible with human values, then the benefits of AI could be transformative - raising the living standards of everyone on earth, ironing out natural human bias and benefitting services in education, medicine, transport, the arts, law, finance and politics, changing the dynamic of history.

However, if we fail in this dauntingly difficult design task, we risk many laws of unintended consequences, from bots designed to kill human targets by sight to the domestic robot cooking the family cat for dinner.

**“The most important book I have read in quite some time.” – Daniel Kahneman**

* “If your privacy alarm bells aren’t ringing… you haven’t been keeping up with the news” p70
* Stuart proposes an updated Declaration of Human Rights: “Everyone should have the right to mental security – the right to live in a largely true information environment” – p 107
* Being concerned about the risks of AI is not the same as being anti-AI. “It deserves a public debate of the highest quality” p 146
* “I expect we’ll be surprised by the range and fluency of machine behaviours… we may be shocked by how sensible they become” p 248
* “the tech industry is going to have to acknowledge that its products matter…I have no doubt that the transition will be a painful one” p252

**ASTONISHING QUOTES**

‘A must-read: this intellectual tour-de-force by one of AI’s true pioneers not only explains the risks of ever more powerful artificial intelligence in a captivating and persuasive way, but also proposes a concrete and promising solution’– **Max Tegmark**

‘The book we’ve all been waiting for. Human Compatible is a brilliant and utterly accessible guide to what will be either the best or worst technological development in human history’ – **Sam Harris**

‘Russell is a leading authority on AI. His new book will educate the public about AI more than any book I can think of ’ **– Judea Pearl**

‘In clear and compelling language, Stuart Russell describes the huge potential benefits of Artificial Intelligence, as well as the hazards and ethical challenges’ – **Professor Martin Re**es

‘Persuasively argued and lucidly imagined, Human Compatible offers an unflinching, incisive look at what awaits us in the decades ahead. Anyone who takes the future seriously should pay attention’ – **Brian Christian, author of *Algorithms to Live By***

‘A totally readable and crucially important guide to the future from one of the world’s leading experts’ – **Tabitha Goldstaub, Head of the UK Government’s AI Council**

‘Stuart Russell has written the book that most – but perhaps not all – machines would like you to read. – **George Dyson, author of *Turing’s Cathedral***

‘This beautifully written book addresses a fundamental challenge for humanity. Essential reading if you care about our future’ **– Professor Yoshua Bengio**

‘A book that charts humanity’s quest to understand intelligence, pinpoints why it became unsafe, and shows how to course-correct if we want to survive as a species’ – **Jaan Tallinn, co-founder of Skype**

**Human Compatible** is the first popular book on AI written by an actual researcher in the field, the result of 40 plus years of work on AI and five years of cross-disciplinary research specifically on how to cope with machines that will eventually be more intelligent than us. Lauded as a thought-leader by the giants of global innovation and tech who themselves are increasingly worried about the risks of AI, he is author of the bestselling text book on AI which is taught in over 1300 universities and is credited with the resurgence of interest in the subject. His 2017 [TED talk](https://www.ted.com/talks/stuart_russell_how_ai_might_make_us_better_people) ‘3 Principles for creating safer AI’ has been viewed over 1.4m times. His appearance in [Do You Trust This Computer?](http://doyoutrustthiscomputer.org/) (Papercut Films, 2018) has been featured in hundreds of media articles. Russell’s Viral Award-winning short film [Slaughterbots](https://www.youtube.com/watch?v=9CO6M2HsoIA) dramatizes the threat posed by lethal autonomous weapons, which has been called by CNN "the most nightmarish, dystopian film of 2017".

**Biography:** Stuart Russell is a professor of Computer Science and holder of the Smith-Zadeh Chair in Engineering at the University of California, Berkeley, where he also directs the Center for Human Compatible Artificial Intelligence. He is an Honorary Fellow of Wadham College, University of Oxford and the vice-chair of the World Economic Forum's Council on AI and Robotics. His work for the UN building a new global seismic monitoring system for the Comprehensive Nuclear-Test-Ban Treaty has been recognized by the Feigenbaum Prize of the Association for the Advancement of Artificial Intelligence. He has been an invited speaker at TED, the World Economic Forum, and the Nobel Dialogues in Stockholm and Tokyo. He is the author (with Peter Norvig) of *Artificial Intelligence: A Modern Approach*, the number one bestselling textbook in AI which is used in over 1,400 universities in 128 countries around the world. He was born in England and lives in Berkeley and Paris.

**Media Contact:** Annabel Huxley, [ahuxley@penguinrandomhouse.co.uk](mailto:ahuxley@penguinrandomhouse.co.uk) tel: 020 7 139 3419