part of the modern journalist's trade. In his acknowledgements, Urbina thanks his front-of-house collaborators, including the journalists, fixers, photographers and videographers who elicited the personal stories that make the statistics credible, as well as the backstage sponsors, more than forty of them, who provided 'material support'.

A clue to Urbina's methods is provided in the racy first chapter, which covers the hunt for *Thunder*. Urbina tells us that he boarded Sea Shepherd's ships in 'early April' 2015 and left them 'a couple [of] days' before *Thunder* was scuttled off the coast of São Tomé on 6 April. That means that the information for his account of the 110-day chase, right down to changes in the weather and the actions of the crews, was scraped from logbook entries and the interviews he conducted with the ships' personnel during his brief period on board.

Are such displays of imagination enough to make this book the last word in gritty studies of the state of the oceans? Probably not, but it's a very good try.

To order this book from the Literary Review Bookshop, see opposite.

JOHN NAUGHTON

Computer Says Go

Human Compatible: Artificial Intelligence and the Problem of Control
By Stuart Russell
(Allen Lane 336pp £25)

The biggest question facing us today in relation to artificial intelligence (AI) is: what if we actually succeed in building superintelligent machines? In particular, what would be the consequences for humankind? This possibility is one of the four 'existential risks' that Martin Rees and his colleagues at Cambridge University's Centre for the Study of Existential Risk are pondering. Such questions go back a long way – at least to 1965, when one of Alan Turing's colleagues, the mathematician I J Good, observed that 'the first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control'.

That proviso about control provides the spur (and the subtitle) for Stuart Russell's book. We are currently living through an intellectual feeding frenzy when it comes to AI, stimulated largely by recent advances in machine learning and its widespread utilisation by the technology giants, together with the spectacular success of machines like DeepMind's AlphaGo, which defeated the world's leading human player of the most difficult known board game, Go.

Because most AI evangelism emerges from industry, the existential risks posed by the technology are often underplayed by its spokespeople. One leading expert, Andrew Ng, for example, famously declared that fearing a rise of killer robots is 'like worrying about overpopulation on Mars'. One possible reason for all this pooh-poohing may be concern that focusing on the risk of AI to humanity might frighten the horses and let on to regulators what tech companies are up to.

Russell is one of those who do not downplay the existential risks. As a distinguished expert in the field he deserves a serious hearing. The essence of his case is first that the arrival of superintelligent machines could indeed be a catastrophe for humanity because they would be indifferent to human needs, preferences and values, and second that our current approach to AI makes the evolution of such sociopathic machines more likely, or even inevitable. What we must therefore do is rethink our approach to AI in order to create machines that complement and enhance humans rather than replace us. We need AI that is, as the title of the book says, human compatible.

Getting there will be a formidable task. Given the momentum that has built up behind the current 'standard model' of AI, changing its direction is the intellectual equivalent of nudging a supertanker onto a radically different course. The first step, Russell argues, is to rethink the concepts of intelligence and rationality that have shaped our attempts to build intelligent machines. We tend to believe that machines are intelligent if their actions can be expected to achieve their objectives. To date, there have been some spectacular successes, such as AlphaGo's mastering of the rules of Go. However, the advance of AI has also had deleterious consequences: Facebook's and Google's machine-learning algorithms for increasing user engagement are undermining democracy. Continuing on this track is not the way to go. As far back as 1960, Norbert Wiener wrote that 'if we use, to achieve our purposes, a mechanical agency with whose operation we cannot efficiently interfere ... we had better be quite sure that the purpose put into the machine is the purpose which we really desire'.

The problem with our current trajectory, Russell argues, is that while we are learning how to design machines that are good at achieving their objectives, we have no reliable way of ensuring that their objectives are the same as our objectives. If we can find a way of fixing that, then superintelligent machines will be beneficial rather than posing an existential threat.

The bulk of the book is devoted to explaining how we might get to being able to design beneficial AI. Because it's a complex and challenging task, and because Russell is an assiduous and conscientious scholar, *Human Compatible* is a long and sometimes demanding read. But if you want to learn something about quantum computing, game theory, probabilistic reasoning, theorem proving, multi-agent negotiations and other relevant but arcane topics, Russell provides a wealth of information.

Admirably, Russell doesn't dodge the obstacles that have to be overcome if we are to design beneficial machines that will augment humanity rather than undermine or subjugate it. But there are times when the lay reader, enmeshed in the thickets of some specialism or other, has to keep a firm grip on where the main argument is heading. This is one of those intellectual voyages where both the journey and the destination matter.

To order this book from the Literary Review Bookshop, see opposite.