TEPSA
Trans European Policy Studies Association

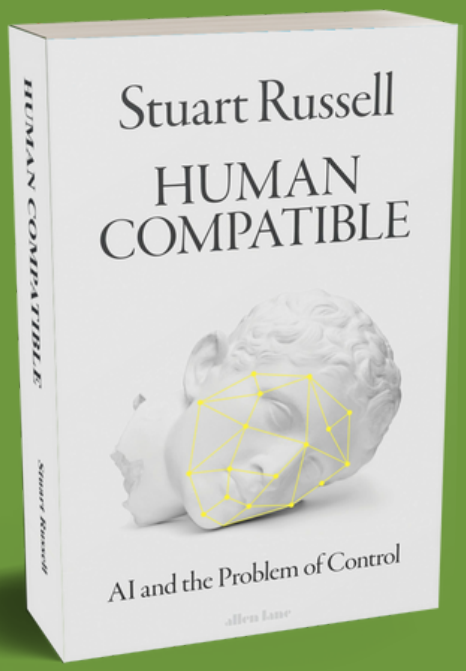
A READING LIST
TO STIMULATE DEBATE

JIM'S READING CORNER

HUMAN COMPATIBLE

AI AND THE PROBLEM OF CONTROL

BY STUART RUSSELL



A fascinating book about AI, its opportunities, and its dangers; it can become the best of all things or the worst of all things. As Russell expresses it: "*The question to ask is 'What if we succeed? Success would be the biggest event in human history... and perhaps the last event in human history.*" Super-intelligent machines can take matters into their own hands, if badly programmed, and wreak havoc. It is possible to manage things in a way to avoid this happening, but it requires careful thinking and a lot of determination.

Russell uses relatively understandable language even though the non-initiated reader will struggle in many places. The book is structured in three parts. The first looks at the very notion of intelligence in humans and in machines. The second sets out the risks that are involved in imbuing machines with intelligence. The last part maps out a new approach to ensure that intelligent machines remain beneficial to humans.

We cannot predict how quickly things will develop. Russell recalls what Lord Rutherford, one of the lost eminent nuclear physicists of the time, said on 11 September 1933 about the possible use of nuclear energy: "*Anyone who looks for a source of power in the transformation of the atoms is talking moonshine.*" The very next day, Leo Szilard, who read about the conference in his newspaper, went for a walk and imagined the neutron-induced nuclear chain reaction.

The concept of intelligence is central for AI. But what do we mean by it? One could say that we are intelligent to the extent that our actions can be expected to achieve our objectives. But what are our objectives? And what are the objectives of the machines we program? The purpose we put into the machine must be what we humans really desire. The objectives of humans and machines must dovetail. The standard model for now of AI is for machines to optimize a fixed objective supplied by humans. Here things look straightforward. But the next steps are about machines learning from errors and making corrections and evolving towards ever greater intelligence.

Another way of looking at the issue is using the notion of *utility*, an invisible property defined by Bernoulli in the 18th century. It is inferred from the preferences exhibited by an individual. In the 1950s John von Neuman developed an axiomatic basis for

a utility theory. A rational agent acts to maximize expected utility. It becomes of course more complicated once more than one person is involved. In this case researchers resort to game theories to better evaluate group utility. Humans are born with their 'agent program' which helps them learn over time to act reasonably successfully across a huge range of tasks. That is not yet the case for AI; machines don't have a general-purpose program; you need to build different types of agent programs. But over time there will be knowledge-based systems, based either on a Boolean logic (with *And* & *Not* gates) or, preferably, on first-order logic; this will be a step towards general purpose intelligence in machines. It is impossible to predict when we will have super intelligent computers: there is a long history of wrong predictions, there is no clear threshold, and AI is inherently unpredictable. The trick to getting there is not better hardware and faster machines. If you are not careful, the latter just gives you the wrong answer faster! To succeed, one needs conceptual breakthroughs: integration of language and common sense/cumulative learning/discovery actions with different hierarchies.

After this first more theoretical part, Russell looks at possible misuses of AI. He distinguishes four categories:

- Surveillance and control (automated STASI/controlling behaviour/taking away mental security),
- Lethal autonomous weapon systems,
- Elimination of work,
- Using human roles.

More generally we face the 'gorilla problem', i.e., the risk of being overtaken by more intelligent machines. (See Samuel Butler's *Erewhon* in 1872 and Frank Herbert's *Dune*). These issues will not go away with denial, deflection, simplistic 'solutions' like 'switching off the machines.

The last part of the book is about possible solutions in the shape of a different approach based on the creation of beneficial machines[1].

How do you define those? In three ways:

- Purely altruistic machines that maximize the realization of human preferences,
- Humble machines that accept a degree of uncertainty as to human preferences,
- Learning machines that learn to predict human preferences from human behaviour.

Provably beneficial AI requires mathematical guarantees, learning from behaviour, assistance games, requests and instructions, learning spurred by internal award schemes, recursive self-improvement. There are many complicating factors before we get to this point: different humans, many humans, nice and nasty humans, stupid and emotional humans; in other words, as Russell tells "*Us*".

At the same time, humans will have to develop adequate AI governance and regulation and action against misuse.

I am not sure I understood everything I report in this short review (that is an understatement). What I do know is that AI is here to stay and will develop exponentially at some point, with huge consequences for the future of mankind. And that these developments will represent a daunting challenge for politicians and authorities that have the difficult task of framing the debate and providing both guidance and regulation and control. Final comment: I strongly believe that the EU has a key role to play in managing AI.

Jim Cloos, TEPSA Secretary-General

[1] In this context it is worth recalling Asimov's three laws of robotics formulated as far back as 1947: the first law is that a robot shall not harm a human, or by inaction allow a human to come to harm. The second law is that a robot shall obey any instruction given to it by a human, and the third law is that a robot shall avoid actions or situations that could cause it to come to harm itself.