**The sci-fi take on robots and the future of work**

Notes for the Diplo event[1]: "Science fiction meets policy | Policy meets science fiction"

Richard Hill[2], 15 January 2019

In Science Fiction, there are at least three paradigms regarding how robots[3] and Artificial Intelligence (AI) [4] will affect humanity, and thus the future of work:

1. Dystopia[5]
2. Utopia[6]
3. Something in between

As we will see below, when looked at in detail, the paradigms are either pessimistic or, in my view, unrealistic and suggest that humanity is not going to be much better off thanks to robots: in the short-run we might recreate an economy like that of ancient Rome, where a privileged few had lots of leisure time and wealth was distributed very unequally; in the long-run, humanity as we know it cannot exist if essentially all work is performed by robots.

**0. Some basics**

Before turning to those three paradigms, it's worth recalling some generally accepted basics about humans. It is generally accepted that humans have evolved from other organisms. The mechanism that drives this is survival of the fittest **gene** (the emphasis is important, see Dawkins' *The Selfish Gene*[7]). That mechanism can be used to explain certain characteristics of humans, e.g. the drive to expand and the need to cooperate (e.g. work in groups). (However, this is not the only possible explanation and it has been disputed).

In much of Science Fiction robots are anthropomorphic[8], not just in form (two legs, two eyes, etc.) but also in behaviour (except that they are usually posited to lack emotions). That is, there is an implicit assumption that robots would have some of the same characteristics as humans, in particular the drive to expand (which assumes that that particular drive is not related to emotions).

But there is no particular reason to suppose that a machine, even one based machine on AI, would have the same evolutionary drive as do humans, unless it were specifically programmed to have it, or unless it acquires it from some external influence (which is what a monolith[9] seems to have done to HAL[10] in the film *2001: A Space Odyssey*[11]). [Semi-scientific aside: as Katharina Hoene pointed out to me, the anthropomorphic tendencies mentioned in the previous paragraph could arise precisely because they are programmed by humans (implicitly taking on of human tendencies) and/or programmed to replace humans (e.g. caring tasks of social robots, explicitly taking on of human tendencies). An assumption underlying AI may be that it will be important to understand and mimic the human brain to go from

---

[1] https://www.diplomacy.edu/calendar/sci-fi-meets-policy-2019
[2] info@apig.ch
[3] https://en.wikipedia.org/wiki/Robot
[4] https://en.wikipedia.org/wiki/Artificial_intelligence
[5] https://en.wikipedia.org/wiki/Dystopia
[6] https://en.wikipedia.org/wiki/Utopia
[7] https://en.wikipedia.org/wiki/The_Selfish_Gene
[8] https://en.wikipedia.org/wiki/Anthropomorphism
[9] https://en.wikipedia.org/wiki/Monolith_(Space_Odyssey)
[10] https://en.wikipedia.org/wiki/HAL_9000
[11] https://en.wikipedia.org/wiki/2001:_A_Space_Odyssey_(film)

primitive to advanced AI. Could a will to survive (even if only on the level of gene) arise implicitly from this quest to mimic human brain structures for advanced AI?]

Alternatively, one might posit that modern AI, which is based on self-learning, might acquire that evolutionary drive by learning from humans and/or nature in general. In such a case, would the AI learn to act emotionally? And would its emotions be distinguishable from those of a human? (This is partly explored in *Ex Machina, Real Humans*, and *Westworld* TV, see references below).

**1. Dystopia**

In the dystopian scenario, either we humans disappear as a species, or we don't allow machines to take over more and more of our work.

There are many variations of this, but, basically, sophisticated machines/robots/androids[12] become sentient/self-aware[13], rebel against humans, and attempt to enslave or exterminate humans. The attempt is more-or-less successful for a while, but is in the end usually defeated. Here are some references:

- R.U.R (1920) play https://en.wikipedia.org/wiki/R.U.R.
- Colossus (1966-1977), novel and film https://en.wikipedia.org/wiki/Colossus_(novel)
- 2001: A Space Odyssey (1968) film https://en.wikipedia.org/wiki/2001:_A_Space_Odyssey_(film)
- Westworld (1973) film and (2016-2018) TV https://en.wikipedia.org/wiki/Westworld_(film) and https://en.wikipedia.org/wiki/Westworld_(TV_series)
- Alien (1978-2017) films https://en.wikipedia.org/wiki/Alien_(franchise)
- Blade Runner[14] (1982, 2017) https://en.wikipedia.org/wiki/Blade_Runner
- Terminator (1984-?) films https://en.wikipedia.org/wiki/Terminator_(franchise)
- Matrix (1999-2003) films https://en.wikipedia.org/wiki/The_Matrix_(franchise)
- Stargate SG1 Replicators (1999-2008) TV https://en.wikipedia.org/wiki/Replicator_(Stargate)
- Real Humans (2012) and Humans (2015) TV https://en.wikipedia.org/wiki/Real_Humans and https://en.wikipedia.org/wiki/Humans_(TV_series)
- Ex Machina (2014) film https://en.wikipedia.org/wiki/Ex_Machina_(film)

The common feature of this view of the future of work is that robots progressively replace humans for tasks that humans don't like to do or that robots can do better. That turns out to be most everything. The robots develop self-consciousness and rebel against their slave-like status.

While it may seem natural to us that sentient beings would rebel against slavery, the historical record is considerably less clear about that. Humans appear to be willing to accept slavery[15] or other forms of subservient status[16] in many cultures that have lasted long periods of time.

From a scientific point of view, it is not clear why machines that were programmed to be subservient to humans would rise against them, nor why machines would have the "selfish gene" evolutionary mechanism that would drive them to evolve in a competitive way. Unless they self-learn to do that. But,

---

[12] https://en.wikipedia.org/wiki/Android_(robot)
[13] https://en.wikipedia.org/wiki/Sentience
[14] This film was based on Dick's *Do Androids Dream of Electric Sheep* (1968) novel https://en.wikipedia.org/wiki/Do_Androids_Dream_of_Electric_Sheep%3F
[15] https://en.wikipedia.org/wiki/Slavery
[16] https://en.wikipedia.org/wiki/Tenant_farmer ; https://en.wikipedia.org/wiki/Domestic_worker ; https://en.wikipedia.org/wiki/Workforce

as far as I can tell, it is only *Ex Machina* (and maybe *Westworld* TV) that implies that self-learning is what caused the robots to rebel against people. Of course self-learning AI is a fairly recent development, so there is no reason why older Science Fiction works should have taken that into account. It will be interesting to see whether future Science Fiction stories account more explicitly for the self-learning aspects of modern AI (which includes perpetuating historical biases found in the training data set[17]).

In the dystopian view, humanity is either exterminated, or reduced to slavery, or manages to defeat the machines/robots/androids and to eliminate them, and to live happily ever after without them. That is, humanity returns to a world in which most work is done by humans, and automation is very limited.

**2. Utopia**

In the utopian scenario, machines create a world of unlimited prosperity, where everyone can have anything they want (except seats at a live event such as a concert, but those are relatively infrequent). People spend their time on whatever leisure activities suit them best: learning, creating, extreme sports, travelling, etc. The most evolved machines manage all this, and can be thought of as benevolent gods.

I've only found one good reference to this:

- Culture series (1987-2012) novels https://en.wikipedia.org/wiki/The_Culture_(series)

It is perhaps telling that the author posits that non-humans (albeit creatures with two legs and two arms) created this world. He may be implying that machines created by humans would not be so benevolent. And it is telling that the gods worshipped by humans are mostly not as benevolent as the *Minds*[18] that govern the *The Culture*[19]. Do the *Minds* mimic the non-humans that created them, or have they somehow transcended them? In our context, can robots/AI be "better" than we are?

In *The Culture*, Machines have no interest in rebelling, or taking power, because (apart from being benevolent) they are citizens with full rights and can vote on important issues (except presumably for *Minds*, which organize votes).

Regarding a world where there is no scarcity of basic resources, animal models might be seals/sea lions in certain areas, lions, and perhaps birds of paradise. Seals of both genders, and male lions, spend very little time hunting and have lots of leisure time. Birds of paradise have evolved very complex mating rituals which also implies lots of spare time. But of course seals and lions are not benevolent as are the humanoids in *The Culture*: they compete fiercely for territory and females (again, this may be explained by "the selfish gene").

**3. In between**

Robots that are useful, but do not threaten humanity or its survival (even if they can be used as soldiers or weapons) appear in many Science Fiction written stories and films. In fact, essentially all written stories and films posit that some things are automated (e.g. autopilots for ships and ground vehicles) or that automation has been banned for good reason (e.g. in *Dune*[20]). Typical depictions of useful robots, and robots as soldiers, are found in *Star Wars*[21]; another notable example is *Star Trek*'s Commander

---

[17] https://www.forbes.com/sites/jasonbloomberg/2018/08/13/bias-is-ais-achilles-heel-heres-how-to-fix-it/#6096b0566e68
https://internetofbusiness.com/mit-researchers-show-how-ai-systems-can-be-made-less-biased/

[18] https://en.wikipedia.org/wiki/Mind_(The_Culture)

[19] https://en.wikipedia.org/wiki/The_Culture

[20] https://en.wikipedia.org/wiki/Dune_(novel)

[21] https://en.wikipedia.org/wiki/Star_Wars

Data[22]. For the purposes of our discussion, we can consider such depictions as trivial and not discuss them further.

[Learned aside: Katharina Hoene has pointed out to me that the full treatment of Commander Data is more complex: he appears to be on a quest to become more human or questions his legal rights (e.g. is he property and is his offspring property); his brother Lore[23] (an AI created before him) is an evil version of Data. The episodes involving Lore imply that he was given a fuller human experience than Data (especially emotions) and that this is the reason why he is evil. This is depicted as a "mistake" that Dr Sung, the creator of both, tried to avoid with Data. This appears to me to be one of the relatively few instances in which Science Fiction has speculated on how emotions could affect robots (other instances being *Westworld*, *Real Humans*, and *Ex Machina*).]

As far as I can tell, the only author who has extensively explored non-trivial paradigms that are in between the dystopian and utopian described above is Isaac Asimov[24].

As shown below, while Asimov's paradigm might seem better than the dystopian paradigm, it actually posits that robots will not be widely used (so it reverts back to the final conclusion of many dystopian paradigms), and that humanity thrives only thanks to robots with certain telepathic[25] abilities (which is even less realistic than *The Culture*'s utopian paradigm).

Asimov's focus on an alternative to the dystopian paradigm is no accident. As Asimov himself explains in the 1983 introduction to *The Naked Sun*[26], he read *R.U.R.* when he was young and started writing about robots in 1939, with the intent to explore non-dystopian paradigms. As part of that exploration, he came up with the three laws[27] of robotics (and apparently coined the term robotics[28]):

> **First Law:** A robot may not injure a human being or, through inaction, allow a human being to come to harm.
>
> **Second Law:** A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
>
> **Third law:** A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

Those laws are so ingrained in the technology used to construct robots (and androids) that they cannot be overridden.

So the dystopian scenario is not possible. Space is conquered by humans who rely on robots, while the people who stay on Earth reject robots and don't live as well, in terms of material comfort.

But a world in which there are huge numbers of robots to do what people would normally do turns out not to be so rosy: the birth rate drops, people become antisocial (to the point of meeting in person only to procreate[29] – or even becoming hermaphrodites so that they don't have to meet to procreate[30]), they

---

[22] https://en.wikipedia.org/wiki/Data_(Star_Trek)
[23] https://en.wikipedia.org/wiki/List_of_Star_Trek:_The_Next_Generation_characters#Lore
[24] https://en.wikipedia.org/wiki/Isaac_Asimov
[25] https://en.wikipedia.org/wiki/Telepathy
[26] https://en.wikipedia.org/wiki/The_Naked_Sun
[27] https://en.wikipedia.org/wiki/Three_Laws_of_Robotics
[28] https://en.wikipedia.org/wiki/Robotics#Etymology
[29] https://en.wikipedia.org/wiki/The_Naked_Sun
[30] https://en.wikipedia.org/wiki/Foundation_and_Earth#Part_IV:_Solaria

lose their ability to cooperate with others (that is, to work in groups)[31], and lose the drive to expand[32]. So in the end the galaxy is colonized by people who don't rely on robots.[33]

Asimov also posits that androids won't be socially acceptable, because of course androids can have sex with humans, and what human (of either gender) would be comfortable competing against attractive androids of the other gender? (The consequences of the sexual attractiveness of androids is also explored *in Real Humans*, *Westworld* TV, and *Ex Machina*).

Here are some references to Asimov's robot stories:

- I, Robot (1940-1950) short stories https://en.wikipedia.org/wiki/I,_Robot
- The Caves of Steel (1954) novel https://en.wikipedia.org/wiki/The_Caves_of_Steel
- The Naked Sun (1957) novel https://en.wikipedia.org/wiki/The_Naked_Sun
- The Robots of Dawn (1983) novel https://en.wikipedia.org/wiki/The_Robots_of_Dawn
- Robots and Empire (1985) novel https://en.wikipedia.org/wiki/Robots_and_Empire

Asimov also posits that humanity is not actually able to make reasonable decisions on its own. So the avoidance of self-destruction and the expansion into the galaxy are only made possible because of two other factors.

First, it turns out that there is a more fundamental law of robotics:

> **Zeroth Law**: A robot may not harm humanity, or, by inaction, allow humanity to come to harm.

Second, and more important, a particular robot has become able to detect and manipulate human emotions, and thus to influence the decisions that humans make. This robot is able to transmit that ability to other robots and androids (created stealthily: people don't realize that they are androids).

So, because of the Zeroth Law, and thanks to their abilities to influence human emotions, the androids manage humans in such a way that they are able to expand, colonize the galaxy, and survive the fall of the Galactic Empire. Here are some references:

- Prelude to Foundation (1988) novel https://en.wikipedia.org/wiki/Prelude_to_Foundation
- Forward the Foundation (1993) novel https://en.wikipedia.org/wiki/Forward_the_Foundation
- Foundation (1951-1953) trilogy https://en.wikipedia.org/wiki/Foundation_series#Foundation_trilogy
- Foundation's Edge (1982) novel https://en.wikipedia.org/wiki/Foundation%27s_Edge
- Foundation and Earth (1986) novel https://en.wikipedia.org/wiki/Foundation_and_Earth

In some ways Asimov's telepathic robots are benevolent gods (and thus not anthropomorphic); they appear to take more decisions on their own than do the *Minds* of *The Culture* and, unlike the *Minds* – with whom any citizen can converse freely – they operate stealthily, without revealing themselves to humans (a characteristic shared with some gods).

It is worth noting that Asimov's robots cannot override their programming, so they are not self-learning in the sense of modern AI. The Zeroth Law turns out to be a consequence of the Three Laws. The

---

[31] https://en.wikipedia.org/wiki/The_Naked_Sun

[32] https://en.wikipedia.org/wiki/The_Caves_of_Steel

[33] https://en.wikipedia.org/wiki/Robots_and_Empire

telepathic abilities were programmed into them (at first by mistake, later deliberately by the robot that first had them, and then by its successors).

Is it possible to program the three laws (or some equivalent) into self-learning AI systems? And would that lead to a fourth law? That is, would self-learning AI systems be benevolent, or would they learn the "selfish gene" characteristics of humans and acquire a drive to expand?

Again, will future Science Fiction explore the paradigms that arise with machine learning, which is a quite different form of AI than the pre-programmed robots that have features in most Science Fiction to date?

**4. Additional references**

Here are some additional references:

- Cognitive Science Movie Index https://www.indiana.edu/~cogfilms/index.php
- https://en.wikipedia.org/wiki/Artificial_intelligence_in_fiction
- https://www.wired.com/story/future-of-work-sci-fi-issue/ short stories, mostly dystopian
- https://en.wikipedia.org/wiki/Black_Mirror

And not related to Science Fiction:

- https://en.wikipedia.org/wiki/Existential_risk_from_artificial_general_intelligence