12 April 2021

Digital Coffee Meeting

Memo by Francisco De Abreu Duarte

Isaac Asimov for policymakers: Beyond the three laws of robotics

Speaker: Nicolas Petit (EUI/Robert Schuman Center)

This event has been organised by the Technological Change and Society Interdisciplinary Research Cluster

Introduction

Science-fiction (s.f.) writers like Isaac Asimov stand out in the policy discussion over Artificial Intelligence (AI) and robotics. Perhaps the most famous s.f. reference in policy circles is Asimov's Three Laws of Robotics.

The three laws are:

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

In 2017, the European Parliament referred to the three laws of robotics in its resolution on Civil Law Rules on Robotics. And in 2020, a member of the French parliament proposed to incorporate the three laws of robotics in the French Constitution.

This prompts the question: what influence should s.f. have on policymaking?

Sci-fi and Policymaking

S.f. works with future facts that can provide important insights for policymaking. Two categories of facts are relevant:

- ✓ Technological facts, i.e. facts about technological change;
- ✓ Social facts, i.e. facts about social responses to technological change.

Although both categories of facts are inherently uncertain, sci-fi can help in building scenarios. Asimov himself explained, s.f. can help "futurism", that is the "respectable specialty thought of by those, in government and industry, who must, every day, make decision by guessing the future". Moreover, there is intuitive appeal to a claim that Asimov's own s.f. might be particularly useful in a policy environment.

There are two reasons: (1) Asimov both wrote s.f. and *on s.f.*; (2) Asimov was himself a scientist with a PhD in chemistry. When he was a student, he worked as of typist for a sociologist. He had a technical grasp of hard science and a professional exposition to social science.

The Asimov Project

The project, led by Nicolas Petit, in joint work with Jerome de Cooman (ULiege), aims at answering two research questions:

- ✓ Are there any high-level lessons to be drawn from Asimov's works?
- ✓ Can the lessons inform policymakers who work on the governance and regulation of AI systems and robotics?

Five take-ways emerge so far, (though these are very preliminary conclusions)

(a) Some technological facts of s.f. literature are more predictive than others. Asimov was very dismissive of a strand of s.f. from the 1960's which disregarded science. He said they are "untrustworthy" as a source of facts. According to Asimov, good s.f. closely tracks the evolution of science.

E.g. Works on time travel = bad sci-fi; predicting the atomic bomb = good sci-fi.

(b) Societies object to technological change, but never discard it completely. Asimov has argued that humans never reject technology completely. Societies tend to fight the social costs generated by technology with some regulation, and *more* technology, not less.

E.g. When robots are somehow banned in Asimov's stories, they are banned from some activities, or sent to other planets to work on tasks relevant to humans. But they continue to play a very relevant role in society.

(c) Societies tend to approach technological risks through a *regulation by design* approach. In most of Asimov's works, the regulatory constraints are coded into the machine.

E.g. The three laws of robotics themselves attest to this. They are coded within the robots' system.

(d) Regulation by design is faulty. The standard policymaking take away from Asimov is that regulation by design is desirable. But Asimov's work was mostly intended to show that regulation by design is fallible, gameable, and did not remove moral hazards. Asimov's ideas (including the three laws) are a cautionary tale for adepts of regulation by design. In later works, Asimov imagined a zeroth law, invented by robots to solve the problems of human made regulation by design

E.g. Imagine two robots. According to the second law, neither of the robots can poison a human. Now, what if the following happens: I ask one robot to fill a glass with a liquid (poison); I ask a second robot to serve the liquid to another human. Neither of the robots is breaching any laws and yet a human was harmed.

(e) Human agency is of the essence. In Asimov's works, there is always an element of human control. One of his leading characters – Susan Calvin – is the prime example of this. Susan is guided by intuition, common sense, and logical experimentation. More than a tech-optimist, **Isaac Asimov is a *tech-institutionalist.*** He believes that code and human agency should complement each other.

Further contributions from the chat/questions:

Q (PhD researcher): Beyond Asimov's inquiry, Philip K. Dick and novels such as *The Minority Report* show how technology can go terribly wrong. Do you think these sci-fi works have a very different outlook on tech regulation? Would we have different solutions according to these authors?

A (Nicolas Petit): Big question. Frankly, I think that Asimov would say that those authors should not have any influence on regulation. They are not predicting the reality; they are not predictive s.f. writers. Asimov would probably describe them as "escape" s.f. He would probably separate them between credible and non-credible.

Q (EUI Professor): I would like to challenge this understanding of Asimov. I think Asimov believes in his laws of robotics even if they fail. He thinks that humanity will fail. One should not forget that in his works, there are super intelligent robots that upgrade the three laws and replace it with a Zero Law (a sort of meta law). There is an idea of humanity as being bigger than any individual protection: robots end up directing History and the future of humanity, leading to experiments and organizing different human societies. It looks like robots adopt a

paternalistic approach to mankind. They are superhuman robots, both in moral and in intelligence, aiming at guiding humanity (as a whole) to the best collective goals possible (even at the expense of individuals).

I would say that the three laws are a kind of constitution for robots - something they cannot avoid following - while the orders would be legislation.

Q (PhD researcher): In the novel 'Runaround', the policy implication is that you cannot resolve problems with laws. Later, in other stories, Asimov hints at the fact that the less is the regulatory oversight, the better the robots will work. In Asimov's novel *The Evitable Conflict*, machines start targeting humans who protest for an anti-robot society. It seems to say that causing small damage to humans to safeguard mankind as a whole is permissible. For Asimov, machines are the only thing which can prevent humans from killing themselves.

Q (Participant): Can technology be independent from human intelligence? What if technology acquires a conscience of its own? How do we tackle this challenge for policy makers? And what about the connection between Asimov and Vinge's technological singularity?

LAST REMARK (Nicolas Petit): Asimov believed only mildly in his laws of robotics, and in his view even the zeroth law was not a panacea. As he wrote, there is an ethical problem enshrined in the Zeroth Law: "to choose between an individual and humanity, when you are not sure of what aspect of humanity you are dealing with, is so difficult that the very validity of Robotic Laws comes to be suspect."[1].

[1] Isaac Asimov, *Robots and Empire* (Doubleday Books, 1985).