Anticipatory science fiction to foster ethical debates on AI and robotics

Carme Torras

Institut de Robòtica i Informàtica Industrial (CSIC-UPC), Barcelona

http://www.iri.upc.edu/people/torras

While in the past the possible social implications of new technological developments could be analyzed before their deployment, now that innovations are constant and become integrated into our daily lives in the blink of an eye, we could say that we are participating in a worldwide experiment without any prior impact study.

The Internet of things, influencers in the network, programs that learn by interacting with humans, assistive and companion robots, computer games with a purpose, serious games for social impact, webs that offer digital immortality… These tools can, in a short time, modify the job market, flip someone's reputation, transform a district, change our relationships —not just at work, but also within our families and our personal relationships— or extend what a person leaves behind after dying, which now includes a digital footprint.

It is difficult to predict —in a substantiated way— the influence that hyperconnectivity and our increasing interaction with machines will have on the evolution of society, the economy, and on people's daily lives. Thus, when trying to anticipate the potential benefits and risks of information technologies, not just lay people resort to science fiction (SF), but some scholars and even companies (e.g., Intel's *The Tomorrow Project*) do so. Quoting the renowned SF writer Neal Stephenson (2011): «Good SF supplies a plausible, fully thought-out picture of an alternate reality in which some sort of compelling innovation has taken place. A good SF universe has a coherence and internal logic that makes sense to scientists and engineers.»

Classical works by Asimov, Dick, or Bradbury already addressed ethically sensible issues, such as those related to mechanical nannies, humanoid replicas, and the regulation of robot development, which have gained traction nowadays due to the upsurge of social robotics. Such upsurge has led to the establishment of a new discipline: Roboethics, a subfield of applied ethics studying both the positive and negative implications of robotics for individuals and society, with a view to inspire the moral design, development and use of so-called intelligent/autonomous robots, and help prevent their misuse against humankind (Veruggio et al. 2011).

Some recent films and TV series tackle also roboethics issues and are used in courses, both online and in the classroom. I would highlight the TV series *Real humans* (where almost human-like robots coexist with people and often compete with them), the film *Surrogates* (in which every citizen has an avatar controlled from home that moves around the city and interacts with people), and the novel *The windup girl* (in which a robot becomes aware that it was built to serve people and wonders about its rights and duties). The film *Robot and Frank* (2012) —showing the relationship between an old man, Frank, and its robotic caregiver— deserves a special mention for its realism and educational value and was the basis for an online course on the *Teach with movies* website, among other places.

Other recent SF works focus on psychological and social issues related to the intensive use of cellphones, widespread interaction in social networks, automatic decision-making based on artificial intelligence, immersive virtual reality games, and learning algorithms using big data, thus triggering interesting debates. In this respect, the TV series *Black Mirror* is a masterpiece that, in each chapter, carries a particular technology to its most extreme consequences, and the movie *Her* depicts a man falling in love with his computer's operating system, thus translating Hoffmann's *The sandman* tale to a digital contemporary setting.

Several universities, particularly in the US, include in their Computer Science and Engineering degrees a course on ethics and human values relative to technology. An opinion shared by professors that have been teaching such a course for several years is that «using fiction to teach ethics allows students to safely discuss and reason about difficult and emotionally charged issues without making the discussion personal» (Burton et al. 2018).

In this context of university education, my novel *The Vestigial Heart* (Torras, 2018) has been published together with online materials to teach a course on *Ethics in Social Robotics and AI*. Six major topics are addressed: how to design the «perfect» assistant, the importance of robot appearance and the simulation of emotions for the acceptance of robots, the role of AI programs in the workplace and in the classroom, the dilemma between automatic decision-making and human freedom and dignity, and civil responsibility versus programmed «morals» in robots. Each topic is developed based on scenes from the novel which tells the story of a teenager, cryopreserved in our time because of an incurable disease and brought back to life in a digital future society. This leads to conflicts with future humans who have been raised by robot nannies, have learned from virtual teachers, and share work and leisure time with AI programs.

Let me conclude with some words that the prestigious journal Nature included in the introduction to the volume entitled *Many Worlds* (2007), commemorating the fiftieth anniversary of the hypothesis of Hugh Everett III about parallel universes, and containing articles from both researchers in quantum mechanics and SF writers. It reads: «Serious science fiction takes science seriously. […] Science fiction does not tell us what the future will bring, but at its best it helps us to understand what the future will feel like, and how we might feel when one way of looking at the world is overtaken by another.» Anticipatory literature has always taken science seriously and has tried to project its accomplishments into the future. It seems that science is also starting to take this literature seriously and find inspiration therein. This confluence could be extremely productive and is very good news, opening up interesting perspectives for the coming years.

References

Burton E., Goldsmith J. and Mattei N. (2018) How to Teach Computer Ethics through Science Fiction. *Communications of the ACM*, 61(8): 54-64. https://cacm.acm.org/magazines/2018/8/229765-how-to-teach-computer-ethics-through-science-fiction

«Many Worlds» (2007) *Nature*, 448(7149): 1-104.

«Robot Ethics» with Robot and Frank (2012) *Teach with Movies website.* http://teachwithmovies.org/robot-and-frank/

Stephenson N. (2011) Innovation Starvation. *Wired*, 27 October. Available at: http://www.wired.com/2011/10/stephenson-innovation-starvation

Torras C. (2018) *The Vestigial Heart. A Novel of the Robot Age*, together with a teacher's guide and a 100-slide presentation to teach a course on *Ethics in Social Robotics and AI*. The MIT Press. https://mitpress.mit.edu/books/vestigial-heart

Veruggio G., Solis, J. and Van der Loos M. (2011) Roboethics: Ethics applied to robotics [from the guest editors]. *IEEE Robotics and Automation Magazine*, 18(1): 21-22.