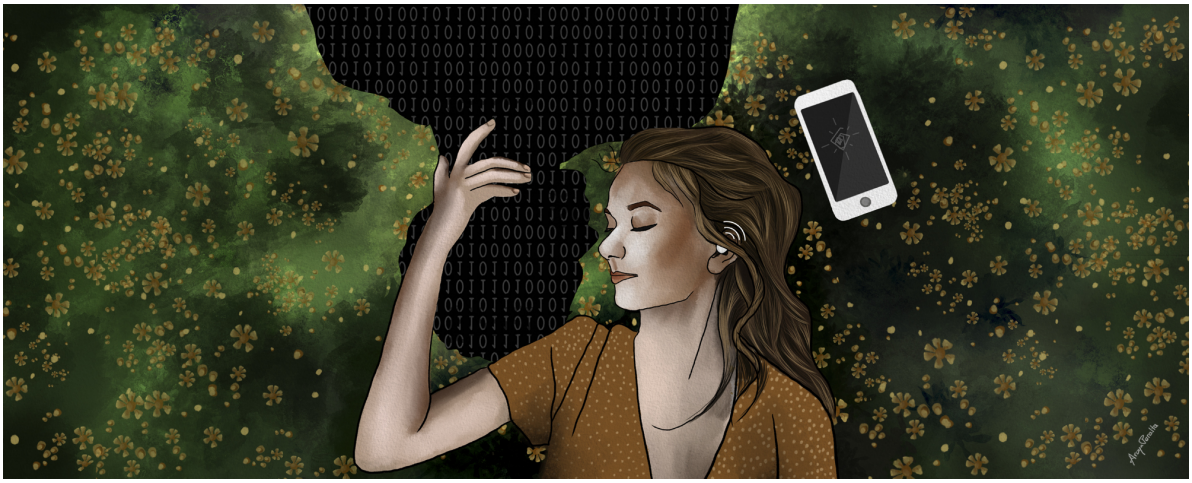
AI'S SYMBOLIC IMAGERIES

Science-Fiction: A Mirror for the Future of Humankind

The engaging debate on AI and ethics favored by science-fiction

Carme Torras



Araya Peralta

Digital technologies have become part of our everyday lives and are increasingly acting as intermediaries in our workplaces and personal relationships or even substituting them. The Internet of things, social networks, programs that learn by interacting with humans, assistive and companion robots, computer games with a purpose, serious games for social impact, roboadvisors, 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 close contacts— or extend what a person leaves behind after dying, which now includes a digital footprint.

The growing interaction with '*intelligent*' machines is not just a further step in the social transformation that started with the industrial revolution. Although these new information technologies do also free humans from repetitive tasks and provide them with more time to spend in creative and enjoyable ways, the difference is that they enter domains previously considered to be exclusive of humans, such as decision-making, emotions and social relationships, which may compromise human values, as well as decisively shape society and our way of life.

This poses a series of ethical questions that were not relevant for other types of machines and about which we have no previous experience, nor can we reliably predict how they will ultimately influence the evolution of humankind. This has led to the confluence of artificial intelligence (AI) with the humanities in an ethical debate that is starting to bear fruit, not only with the establishment of regulations and standards, but also with educational initiatives in university teaching, professional improvement, and the conformation of public opinion. Interestingly, science fiction (SF) often plays a prominent speculative role in highlighting the pros and cons of potential scenarios. Quoting the renowned writer Neal Stephenson [1]: «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».

Confluence of AI with the humanities

The AI research community has felt the need of working together with social scientists, psychologists, lawyers, philosophers and anthropologists, to analyze the social and ethical implications of the technologies being developed and the way they are deployed. Multidisciplinary teams have been formed, joint research programs launched, and governments and institutions have carried out surveys and established committees to get hold of the situation and come up with guidelines and standards.

Regarding the latter, several organizations and professional societies are developing regulations for AI programmers, robot designers, companies and users. In April 2019, the European Commission's High-Level Expert Group on AI presented Ethics guidelines for trustworthy AI, and later that year the IEEE Standards Association published Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, which had been open to public discussion since December 2017 and received over three hundred pages of in-depth feedback that was reviewed by one thousand experts from the realms of business, academia and policy. Previously, there had been other pioneer, more modest initiatives, such as the Barcelona Declaration for the proper development and usage of AI in Europe and the Montréal Declaration for a Responsible Development of Artificial Intelligence, both issued in 2017.

Legal regulation is needed for enforcing Ethical AI, but even more important is education at all levels, so as to develop critical awareness and informed public opinion. A crucial target in this regard are those designing and shaping AI tools. It is not surprising that, at the university level, renowned voices such as that of Prof. Barbara J. Grosz advocate for integrating ethics in computer science education: «By making ethical reasoning a central element in the curriculum, students can learn to think not only about what technology they could create, but also whether they should create that technology.» [2]. Along this line, prestigious associations such as the Association for Computing Machinery (ACM) and the Institute of Electrical and Electronics Engineers (IEEE) include 18 knowledge areas in their Computer Science curricula, one of which is *Social Issues and Professional Practice*, so that «students develop an understanding of the relevant social, ethical, legal and professional issues».

In Europe, the University of Oxford has even gone a step further in the confluence with the humanities by offering a degree in Computer Science and Philosophy, which focuses on common interests in artificial intelligence, logic, robotics, virtual reality, and ethics. Such mixed degrees will soon proliferate all over the world and, in particular, several Catalan universities have initiatives in this direction.

The role of Science Fiction

Science Fiction (SF) has been a source of inspiration for scientists and technologists since its outbreak. Classical works by Shelley, Verne, and Huxley, among many others, envisaged research adventures, discoveries and innovations that later have become true, like space exploration, humanoid replicas, and genetic engineering. Nowadays, SF has gained increasing public attention due to the upsurge of digital technologies; most notably, AI and robotics. Many initiatives to put in contact SF writers and filmmakers with scientists and technology developers have emerged in recent years.

As an example, in 2012 the Center for Science and Imagination was set up at Arizona State University. The idea came from a thought-provoking speech by Neal Stephenson [3], given in the presence of the university president, in which the writer stated that today's scientists had lost the ability to think and do «great things», like those that had previously inspired the Apollo space program or the microprocessor. The president responded that perhaps it was the SF writers who were at fault because they were failing to evoke an ambitious future that would inspire scientists to make them into a reality. As a result, the Center now houses teams of researchers in science and humanities to devise and pursue ambitious goals that shape the future.

One of the projects run by the Center is the continuation of an initiative launched by the company Intel, *The Tomorrow Project*, in which they asked four SF writers to create stories picturing possible future uses of its products in photonics, robotics, telematics and smart sensors. The book with the four accounts is open access [4], and several volumes have appeared since then in which solutions are proposed to the greatest challenges facing humanity today, through the visual arts and writing, all as a result of the work carried out at this Center.

Countless initiatives in this direction have emerged: joint workshops, prospective symposia, exhibitions, performances, and even top scientific journals have edited special issues devoted to SF inspiration. A pioneer was the prestigious journal Nature, which to commemorate the fiftieth anniversary of the hypothesis of Hugh Everett III about parallel universes, published a volume entitled *Many Worlds* [5] containing articles from both researchers in quantum mechanics and SF writers. Its introduction states very clearly what role SF can play in anticipating the benefits and risks of scientific development: «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.»

Teaching and fostering ethics debate using SF stories

Teaching Ethics in AI differs considerably from teaching other subjects in Information Sciences and Engineering. It is not so much a matter of students learning some specific contents, but making them aware of the social and ethic implications of their future jobs and train them to analyze and debate about such issues. People hold multiple and often conflicting sets of values and the aim is not to unify the views of students around a set of rules, but to raise their awareness and abilities to think and discuss. Moreover, technology students are not philosophers. Although there are consolidated ethical theories that they should know about, philosophical texts are often too abstract for them, and a pragmatic option is usually taken.

According to Sullins [6], the main ethical theories relevant to AI and human-robot interaction are: consequentialism or utilitarianism (maximizing the number of people that enjoy the highest beneficial outcomes), deontologism (acting only according to maxims that could become universal laws), virtue ethics (relying on the moral character of virtuous individuals), social justice (all human beings deserve to be treated equally and there must be a firm justification in case of mistreatment), common goods (living in a community places constraints on the individual), religious ethics (norms come from a spiritual authority), and information ethics (policies and codes for governing the creation, organization, dissemination, and use of information).

Since no single theory is appropriate for addressing all ethical issues arising in the design and use of technical innovations, the pragmatic option is to adopt a hybrid approach. Such hybrid ethics is advocated by Wallach and Allen [7] as a combination of top-down theories (i.e., those applying rational principles to derive norms) and bottom-up ones (i.e., those inferring general guidelines from specific situations).

Now, where should these specific situations come from? Stephenson [8] claims: «What SF stories can do better than almost anything else is to provide not just an idea for some specific technical innovation, but also to supply a coherent picture of that innovation being integrated into a society, into an economy, and into people’s lives.» Thus, some Ethics in Technology courses recur to SF stories to exemplify conflictive situations. Themes addressed in the classic works by Asimov, Dick, Bradbury, Orwell, Huxley, Hoffman, Shelley, Capek, Wells, Sturgeon, Silverberg, or Keyes, such as the three laws of robotics, robot nannies, security versus freedom, lack of privacy, technological totalitarism, emotional surrogates, humanoid replicas, incidence on the job market, moral responsibility, loss of human control, high-tech biases, manipulation and automation divides, or human enhancement and posthumanism, have attained great relevance with the development of AI.

Given this relevance, it is natural that instructors teaching Ethics in technological degrees are recurring to such SF stories to exemplify sensitive situations the students may face in their professional practice so as to foster a fruitful reflection and debate about them. After teaching the course *Science Fiction and Computer Ethics* five times at the University of

Kentucky and two times at the University of Illinois at Chicago, Burton et al. [9] state on their experience: «Using fiction to teach ethics allows students to safely discuss and reason about difficult and emotionally charged issues without making the discussion personal.» Besides highlighting how engaging narrative is for students, the authors report on many positive insights they got along the years that are worth reading in detail.

Modern science fiction touches upon many of the ethical issues depicted in classical stories, but usually focuses on the concerns raised nowadays by digital technologies, such as those derived from our 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 [10]. Figure 1 shows some series and movies focusing on these concerns. In this respect, the 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. Related to robotics, I would highlight the series *Real Humans*, where almost human-like robots coexist with people and often compete with them, and the film *Surrogates*, in which every citizen has an avatar controlled from home that moves around the city and interacts with people. The film *Robot and Frank*, showing the relationship between the elder Frank and its robotic caregiver, deserves a special mention for its realism and educational value and was the basis for an online Robot Ethics course on the Teach with Movies website [11], among other places.

Figure 1. Science Fiction series and movies dealing with ethical issues raised by digital technologies, especially AI and robotics.

Turning to books specifically designed to teach courses in technological degrees based on science fiction, I would highlight Murphy [12], Nourbakhsh [13] and Torras [14]; see Figure 2. The one edited by Murphy is intended to explain key principles of AI; the enclosed stories —by Asimov, Vinge, Aldiss, and Dick— cover telepresence, behavior-based robotics, deliberation, testing, human-robot interaction, the *uncanny valley*, natural language understanding, machine learning, and ethics. Each story is preceded by an introductory note, "As You Read the Story," and followed by a discussion of its implications, "After You Have Read the Story."

Figure 2. Three books published by MIT Press specifically designed to teach courses in technological degrees based on science fiction.

The other two books differ from this one in two respects: i) the authors themselves have written the science fiction stories used for illustration, and ii) the books do not explain technical principles, but are explicitly intended to foster debate on the social and ethic implications of information and communication technologies.

By drawing possible future scenarios, Nourbakhsh raises some concerns about where we are heading, without neither taking an ethics regulatory viewpoint nor explicitly trying to be pedagogical. The author, a renowned roboticist, makes very lucid remarks by concentrating the following specific topics: marketing strategies in the net; the consequences of non-ephemeral design; robotic flying toys that operate by means of gaze tracking; robot-enabled multimodal, multicontinental telepresence; and even a way that nanorobots could allow us to assume different physical forms. Nourbakhsh examines the underlying technology and the social consequences of each scenario. He also offers a counter-vision: a robotics designed to create civic and community empowerment.

Because of my research on assistive robotics [15], I became progressively concerned about the social and ethics implications of the technologies we are developing and especially interested in devising ways to teach Ethics to technologists. This led me to try my hand at fiction, and in the novel *The Vestigial Heart* [16] I imagined how being raised by artificial nannies, learning from robotic teachers and sharing work and leisure with AI programs would affect the intellectual, emotional and social habits of future generations. The novel's

leit motiv is a quotation from the philosopher Robert C. Solomon: «it is the relationships that we have constructed which in turn shape us». [17] He meant human relations with our parents, teachers and friends, but the quotation can be applied to robotic assistants and all sorts of interactive devices, if they are to pervade our lives.

The plot of the novel is as follows: Celia, a-thirteen-year-old girl cryogenically frozen because of her terminal illness, is cured and brought back to life in the 21st Century in order to be adopted. Aside from her memories, she brings something else from the past: feelings and emotions that are no longer in existence. These are what most attract Silvana, a middle-aged woman who works as an emotional masseuse trying to recover the sensations humans have lost. Celia's feelings are also precious research material for Leo, a bioengineer who is designing a creativity prosthesis for the mysterious Doctor Craft, owner of the leading robotics company, CraftER.

Following a suggestion by MIT Press Editor Marie L. Lee, an appendix with 24 ethics questions and hints for a discussion around the situations appearing in the novel was included in the book, and published together with an online teacher's guide and a 100-slide presentation to deliver a course on *Ethics in Social Robotics and Artificial Intelligence*. It covers six major topics: 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 educational environments; the dilemma between automatic decision-making and human freedom and dignity; and civil responsibility related to embedding «morals» in programs and robots.

Each section in the teacher's guide follows the same structure, starting with some highlights from the novel, then the corresponding ethics background is provided, followed by four questions and hints for their discussion, and closing with some revisited issues from previous chapters. Slides showing the structure of these ethics teaching materials and some of the questions addressed are displayed in Figure 3.

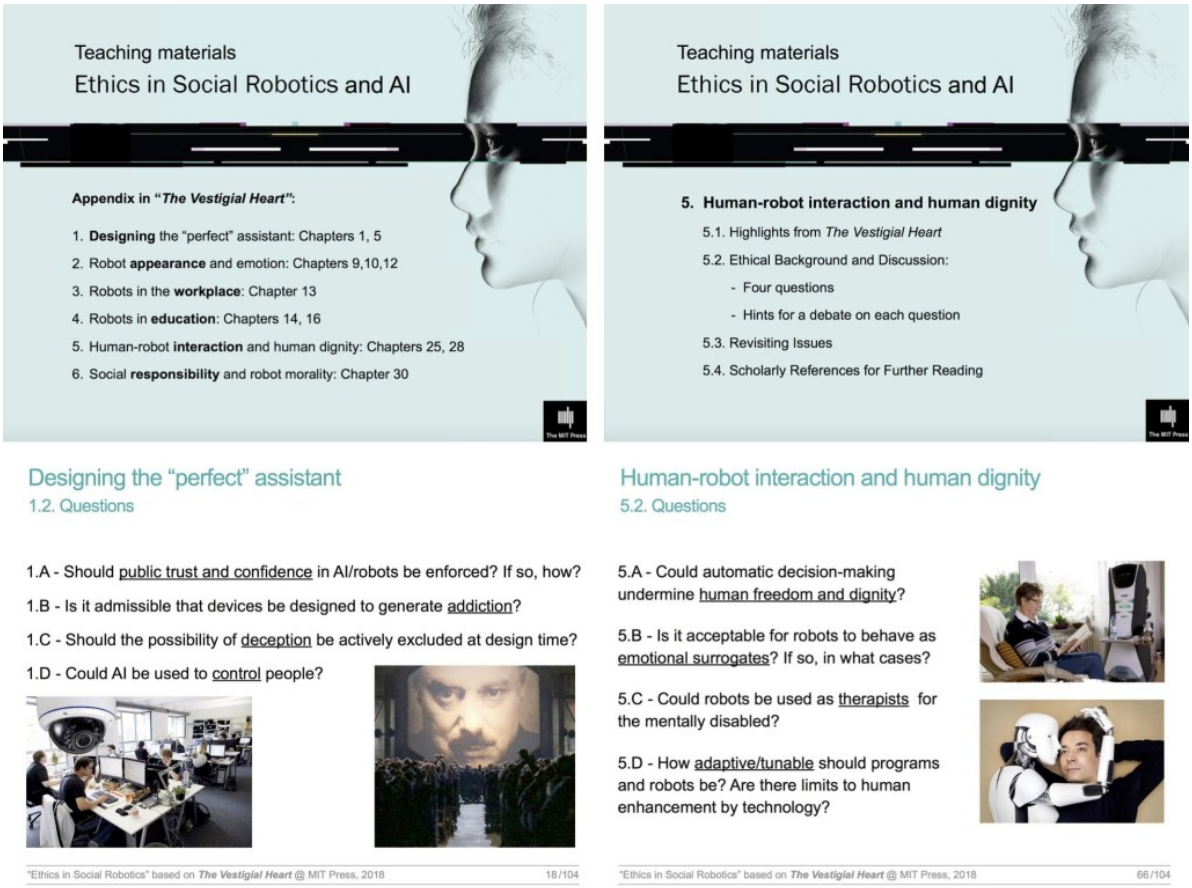


Figure 3. Sample slides from the teaching materials associated with The Vestigial Heart showing: (top) correspondence between ethics sections and chapters of the novel, as well as section structure, and (bottom) some of the questions addressed.

The book together with the ancillary materials have been used not only in Information Sciences and Engineering degrees, but also in Philosophy and Business Administration studies at several universities, mainly in Spain and USA so far, but some European universities have already expressed interest to do so in the coming academic terms. Moreover, a Catalan version of the materials adapted to Secondary School is available at the web of Pagès Editors, which can also be used to trigger debate in reading groups, panels, round tables, and similar discussion forums.

Concluding remarks

The increasing interaction of people with AI programs in everyday life poses important social and ethical challenges with a lot of potential to substantially shape our future. This calls for technical degrees getting closer to the humanities, so that students and professionals become aware of possible sensitive issues they may face in their careers and learn to reflect and discuss about them.

We have described three recent proposals by researchers in AI and robotics: one book using classical stories to explain computational principles, and two books relying on modern science fiction to convey technoethics knowledge, with the singular characteristic that fiction was written by the researchers themselves so that the stories accurately illustrate the issues to be discussed, in addition to making the material pedagogically

appealing to technology students, instructors, and also possibly general readers interested in these amazing future perspectives.

Let me conclude by saying that most SF literature has taken science seriously and has tried to project its accomplishments into the future. It seems that science is also starting to take this anticipatory 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.

ACKNOWLEDGEMENT

This work has been partly supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme through the project CLOTHILDE – CLOTH manIpulation Learning from DEmonstrations (Advanced Grant agreement No 741930), and the Spanish Research Agency through the María de Maeztu Seal of Excellence to IRI (MDM-2016-0656).

REFERENCES

1 — Stephenson, N. (2011) "Innovation starvation". World Policy Journal, 28(3), 11-16. Disponible en línia a: http://www.worldpolicy.org/journal/fall2011/innovation-starvation

2 — Grosz, B.J.; Grant, D.G.; Vredenburgh, K.; Behrends, J.; Hu, L.; Simmons, A. i Waldo, J. (2019) "Embedded EthiCS: integrating ethics across CS education". *Communications of the ACM*, 62(8), 54-61.

3 — Stephenson, N. (2011) "Innovation starvation". World Policy Journal, 28(3), 11-16. Disponible en línia a: http://www.worldpolicy.org/journal/fall2011/innovation-starvation

4 — Rushkoff, D.; Hammond, R.; Thomas, S. i Markus, H. (2012) *The Tomorrow Project*. Els millors escriptors descriuen la vida quotidiana en el món del futur. Intel: Santa Clara.

5 — Everett III, H. (2007) «Many Worlds» *Nature*, 448(7149): 1-104.

6 — Sullins, J.P. (2015) "Applied professional ethics for the reluctant roboticist". Dins de *The Emerging Policy and Ethics of Human-Robot Interaction*, editat per L.D. Riek, W. Hartzog, D. Howard, A. Moon i R. Calo. Taller de treball dins el 10è Congrés Internacional de l'ACM/IEEE sobre la Interacció Humans-Robots, Portland.

7 — Wallach W. i Allen C. (2008) *Moral machines: Teaching robots right from wrong*. Oxford: Oxford University Press.

8 — Stephenson, N. (2011) "Innovation starvation". World Policy Journal, 28(3), 11-16. Disponible en línia a: http://www.worldpolicy.org/journal/fall2011/innovation-starvation

9 — Burton, E.; Goldsmith, J. i Mattei, N. (2018) "How to teach computer ethics through science fiction". *Communications of the ACM*, *61*(8), 54-64.

10 —

• Torras C. (2018) "Social networks and robot companions: Technology, ethics and science fiction". *Metode Science Studies Journal*, 99: 47-53.

• Torras C. i López de Mántaras, R. (ed.) (2019) "Interlinked: machines and humans facing the 10101 century". *Metode Science Studies Journal – Annual Review*, 9.

11 — Teach with Movies (2012) “Robot Ethics Using Clips from *Robot and Frank”*. http://teachwithmovies.org/robot-and-frank/
12 — Murphy, R. (ed.) (2018) *Robotics Through Science Fiction: Artificial Intelligence Explained Through Six Classic Robot Short Stories*. MIT Press: Cambridge, Massachusetts.
13 — Nourbakhsh, I.R. (2013) *Robot Futures*. MIT Press: Cambridge, Massachusetts.
14 — Torras C. (2018) *The Vestigial Heart*. *A Novel of the Robot Age.* MIT Press: Cambridge, Massachusetts (llibre del professor i presentació de 100 diapositives com a material de suport per al docent (en anglès) disponible en línia a: http://mitpress.mit.edu/books/vestigial-heart)
15 —
• Torras C. (2016) “Service robots for citizens of the future”. *European Review*, 24(1), 17-30.
• Torras C. (2019) “Assistive robotics: Research challenges and ethics education initiatives”. *DILEMATA: International Journal of Applied Ethics*, 30: 63-77.

First published in Catalan as La mutació sentimental.

17 — Solomon, R.C. (1977) *The Passions*. Nova York: Anchor Press / Doubleday.

Carme Torras

Carme Torras is a novelist and research professor at the *Institut de Robòtica i Informàtica Industrial* (Institute of robotics and industrial computing), a joint research centre of the Spanish Council for Scientific Research (CSIC) and the Technical University of Catalonia (UPC). She graduated in Mathematics and Computer Science at the Universitat de Barcelona and the University of Massachusetts, respectively, and holds a PhD degree in Computer Science from the Universitat Politècnica de Catalunya. She is IEEE Fellow, EurAI Fellow, member of the European Academia and member of the Reial Acadèmia de Ciències i Arts from Barcelona. She combines her role leading a research group in assistant robotics at the CSIC-UPC with her work as a fiction writer and promoter of ethics in the application of new technologies. She authored several books and articles, as well as science-fiction novels related to robotics and artificial intelligence.