Writing Science Fiction as an inspiration for AI research and Ethics dissemination*

Carme Torras$^{1[0000-0002-2933-398X]}$ and
Luís Gustavo Ludescher$^{2[0000-0001-8915-2032]}$

[1] Institut de Robòtica i Informàtica Industrial, CSIC-UPC
Llorens i Artigas 4-6, 08028-Barcelona, Spain
torras@iri.upc.edu
http://www.iri.upc.edu/people/torras
[2] Department of Computing Science, Umeå University
MIT-huset, 901 87 Umeå, Sweden
luisl@cs.umu.se
https://www.umu.se/en/staff/luis-ludescher/

Abstract. In this chapter we look at science fiction from a perspective that goes beyond pure entertainment. Such literary gender can play an important role in bringing science closer to society by helping to popularize scientific knowledge and discoveries while engaging the public in debates which, in turn, can help direct scientific development towards building a better future for all. Written based on a tutorial given by the first author at ACAI 2021, this chapter addresses, in its first part, how science and science fiction can inspire each other and, in its second part, how science fiction can be used as an educational tool in teaching ethics of AI and robotics. Each of the two parts is supplemented with sections containing the questions asked by the audience during the tutorial as well as the provided answers.

Keywords: Science fiction · Ethics · Artificial intelligence · Robotics · Education · Science and technology · Techno-ethics.

1 Learning objectives

- Familiarize the audience with cases of mutual inspiration between science and fiction, as well as initiatives to promote joint work in multidisciplinary teams.
- Increase awareness of new ethics issues raised by digital technologies, in particular Artificial Intelligence and Robotics.
- Learn about ways in which techno-ethics has been incorporated in engineering and computer science university degrees.

* Supported by the European Research Council (ERC) within the European Union Horizon 2020 Programme under grant agreement ERC–2016–ADG–741930 (CLOTHILDE: CLOTH manIpulation Learning from DEmonstrations)

- Get to know recent science fiction literary works that have been used (or can be used) to teach techno-ethics and trigger debate.
- Providing the attendants with a hands-on experience of the teacher's guide and presentation accompanying the novel *The Vestigial Heart*, so as to enable them to teach a course, some sessions or just trigger debate on "Ethics in AI and Robotics" using these free-of-charge ancillary materials [1].

2 Science and fiction: mutual inspiration

The mutual inspiration between science and fiction is now more apparent than ever, perhaps due to the acceleration of scientific progress in recent decades and the emergence of digital technologies, which seem to have installed us in a fantasy future. Although only a few of us, as William Gibson, the father of Cyberpunk, said in an interview: "The future is here, but it is unevenly distributed." Fiction can also be inspiring in this regard, both by promoting the critical thinking needed to avoid dystopian risks and by guiding scientific research with an ethical perspective that leads us to a more equitable and sustainable future.

Just take a look at today's series, movies and narrative to see the growing presence of science fiction, and how recent scientific discoveries and technological innovations are the seed of most plots. There is no doubt that science is now a great source of inspiration for fiction. Every time science makes a discovery that revolutionizes our view of things, science fiction takes the idea and expresses it to the last detail. The splendid *Black Mirror* series is a good example of this: each chapter takes a digital technology, such as virtual reality, brain implants, social media, learning avatars or online games, and develops it often to the most dramatic extreme. The series is a true masterpiece, with a great ability to raise awareness and generate excellent debate.

And vice versa, there are many examples of science fiction inspiring science, from Jules Verne's "Twenty Thousand Leagues Under the Sea" to "Star Trek" with gadgets like the communicator, the forerunner of the mobile phone, not to mention robots of all kinds that have filled the imagination of past generations and have now become a reality even in the most complex cases such as robots with humanoid physiognomy. The list goes on and on, but perhaps it is worth noting how much space exploration owes to dreamers who imagined what might be on the neighboring planets.

The more we discover about our universe, the more fiction we write; the more fiction we write, the more inspiration there is for scientists. This is the fruitful feedback loop that we would like everyone to share, as adding the ethical ingredient would lead research, technology deployment and innovation to a deeper knowledge not only of what surrounds us, but also of our place in this uncertain future, in which we would like to make utopias possible such as 'evenly distributing' technological resources as suggested by Gibson.

3 Fostering the confluence of technoscientists and writers/artists

Many initiatives to put in contact SF writers and filmmakers with scientists and technology developers have emerged in recent years to foster the aforementioned feedback loop between science and fiction. Besides joint workshops, interdisciplinary research projects, exhibitions and performances, 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 [2] 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 understand what the future will feel like, and how we might feel when one way of looking at the world is overtaken by another.”

Among other similar initiatives, MIT Technology Review publishes annually a volume entitled *Twelve Tomorrows* consisting of science fiction stories about different technologies, and Intel promoted *The Tomorrow Project* compiling four short stories about their products projected into the future. Even MIT Press is recently publishing science fiction stories to teach economics, artificial intelligence, robotics and roboethics.

We can thus conclude that that science is also beginning to take science fiction seriously.

Let us mention another bridge that has been established between science, technology and science fiction. Neal Stephenson, a renowned science fiction writer, gave a talk entitled ”Innovation starvation” [3], where he claimed that scientists had lost imagination after the big challenges of the microprocessor and the space exploration missions, and that they desperately needed imagination from science fiction writers. He said so in front of the Chair of Arizona State University, who took the challenge and created the Center for Science and the Imagination [4], which hosts interdisciplinary teams of artists and scientists to tackle very bold ideas, projects that may not seem feasible today, but they give them a try.

4 New ethics issues raised by Artificial Intelligence

Working on assistive robotics [5], which entails a lot of AI-based human-robot interaction, has made us reflect about ethics and the social implications of the work we do. Many of the issues to be addressed concern programs that learn. As a simple example, some years ago Microsoft placed the Tay chatbot in Twitter to have regular conversations with people and, after only 16 hours, it had to be removed because it had turned nasty, racist, sexist and it insulted everyone. This tells us that our interactions in the Internet can be learned by this type of programs that learn, and we should be very careful, so the responsibility is no longer only at the programmer side, it is also at the users side.

You may also have heard that this type of learning programs are prone to biases. Nowadays deep learning is trained with big data, which may have the biases inherent to historical records. There is a famous case of the judicial system in the United States that used a program to determine the probability of recidivism of prisoners and, since it was trained with historical data, it predicted that Afro Americans would recidive more than white people. This is what statistically happened in the past. Biases are in the data, in the history, in our prejudices, not in the learning algorithms themselves. Therefore, it is very important to clean up big data before training the algorithms.

Many of these issues are new, emerging now due to the spreading of AI applications. Another one is related to webs that offer to remove all our interventions in Internet, such as the web called Suicide Machine [6]. You know this cannot be entirely done: some posts can be at least hidden in platforms like Facebook or Twitter, but all the things that were downloaded by particular users to their computers cannot be removed, thus we should be very careful of what we put in in these social networks and webs. And there is also the opposite: webs that offer virtual immortality. How? By compiling all our interventions in the net; for instance, all the likes we place, the things we buy, the movies, the photographs, who we interact with, our opinions, etc. are there and configure an avatar of us. They offer this avatar, first when you are alive just to answer mail for you, but for people that die, they offer it also to their families, saying that they can continue interacting with the dead relative. This is also plotted in an episode of the series we were mentioning at the beginning, *Black Mirror*. But it happens also in reality, there are many, up to four at least, webs that offer this virtual immortality [7–10]. Things are changing quickly and we must be aware of all this.

Some other issues raised by AI and social robotics were shared with other technologies in the past. For instance, the incidence on the job market is not new, since in the industrial revolution some tasks or jobs were automated and this had implications for people. Likewise, legal liability, privacy and social divides are not new issues. What is new about AI and all these new digital technologies? That these programs enter domains that were thought to be just human, like communication, feelings and relationships, and decision making. Some of these programs, for instance in the medical area, may make decisions for us and this should be regulated. Moreover, they may intervene in our feelings and relationships; we get attached to these devices that help us so much and may even have a human or quasi-human appearance. Implants and human enhancement through technology raise also many questions, and people are debating a lot about bioconservatism, transhumans and posthumans.

In sum, the field of techno-ethics appeared already some years ago because the AI community and the robotics community are really very worried about all these issues, not to mention military applications. Techno-ethics is a subfield of applied ethics studying both the positive and negative implications of robotics and AI for individuals and society. And it has two branches: one is human ethics applied to robotics and AI, where the history of philosophical thinking about ethics can be applied and some conclusions be drawn for these new issues that are appearing. And a second, new branch is that some codes of ethics are being embedded in the robots and programs themselves, what is termed machine ethics. There are many research groups working on values to be introduced in the programs so that there are some red lines that these programs cannot cross.

Let us talk just a bit about organisms, institutions and professional societies that are putting in place regulations and standards for these new issues that are appearing. For instance, the European Parliament launched the ”Civil law rules on robotics”[3] four years ago already, which are very general principles derived from human rights. All regulations we will mention are available on the Internet. Specifically for AI, the high-level expert group on AI issued a draft on ”Ethics guidelines for trustworthy AI”[4]. It is very informative, much more detailed than the aforementioned rules for robotics, and very handy for people working in this area of artificial intelligence.

There are also several networks and AI projects that have also issued some ethics guidelines. The Standards Association within the IEEE, the Institute of Electrical and Electronics Engineers, has released the document ”Ethically Aligned Design”[5] that covers 12 areas in which robots and intelligent systems are being applied, for instance at schools, in the workplaces, the military also, etc. This is a well developed and very informative document, whose reading for specific issues we recommend very much.

5 Introducing techno-ethics in the curricula

Complementary to regulation is ethics education at all levels, from primary school, to high-school, university and the general citizenship. Here we will focus on introducing techno-ethics in the curricula of technological degrees, and especially teaching it through science fiction narrative. Since several years ago, one of the 18 knowledge areas in the IEEE ACM Computer Science curriculum[6] is ”social issues and professional practice”, that includes courses on professional ethics, technology and society, and the like.

Barbara Gross, a professor in computer science at Harvard University, says that ”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” [11], thus making students aware of the implications of creating specific technologies. And this type of courses

[3] https://www.europarl.europa.eu/RegData/etudes/STUD/2016/571379/IPOL_STU(2016)571379_EN.pdf

[4] https://data.europa.eu/doi/10.2759/346720

[5] https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf

[6] https://www.acm.org/binaries/content/assets/education/curricula-recommendations/cc2020.pdf

is taught now not just in the US, but also in Europe and in several universities worldwide. They were usually taught using philosophical textbooks, which proved to be a bit too abstract for technology students, so more and more professors and teachers are using classical science fiction readings in these courses.

6 Science fiction narrative engages technology students

Neal Stephenson, whom we mentioned earlier, advocates for the use of science fiction to anticipate future scenarios with ethics implications, and he says that ”what science fiction 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” [3], therefore drawing scenarios that permit imagining and discussing about the good and bad aspects of these new technologies.

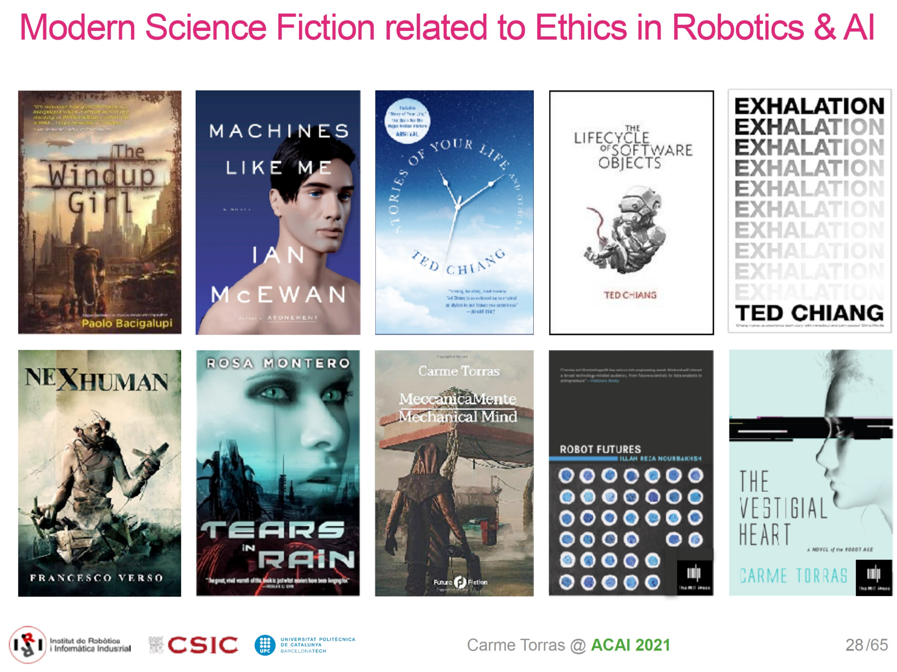
Judy Goldsmith, who is a professor of computer science at Kentucky University, has been teaching a techno-ethics course using science fiction for seven years already, and she collected her experiences in a very nice paper [12]. One of her conclusions is that ”using science fiction to teach ethics allows students to safely discuss and reason about difficult and emotionally charged issues without making the discussion personal”.

Then, what type of classic science fiction stories is she and other professors using? Basically, classical works by Isaac Asimov, that anticipated robot nannies, by Philip K. Dick and Mary Shelley, anticipating androids, humanoid robots, and E.T.A. Hoffman, anticipating robot companions, robot nannies, and so on. But these are works from the last century, and nowadays teachers are progressively using more up-to-date, recent science fiction novels and short stories to exemplify future situations, since those depicted in last-century narrative are almost present, if not already past.

Next we mention some novels that we like and recommend, which raise a lot of questions that can foster debate on techno-ethics (Fig. 1). One is *The Windup Girl* by Paolo Bacigalupi, in which a robot that has been designed to serve, when she (because it's a female robot) is worn out and discarded, becomes aware of her having been designed specifically for this purpose and reflects on the issues this raises for and around her.

Ian McEwan, a very renowned writer, recently wrote the novel *Machines Like Me.* Observe the ambiguity of the title, it can be that machines like the person or that machines are like the person. This book raises some issues that are very up to date; for example, who is allowed to initially configure a robot or a learning application for their own purposes, in this case at home. There are several members in the family and each has their own goals and preferences, some of which might be conflicting. Another question is who is entitled to shut the robot or program down when it does something that the current user or someone else doesn't like.

The short-story writer Ted Chiang has become very famous after publishing just two compilations of stories, namely *The Story of Your Life* and *Exhalation.*



Modern Science Fiction related to Ethics in Robotics & AI

Carme Torras @ ACAI 2021 28/65

Fig. 1. Examples of fiction narrative that can be used to trigger debate on techno-ethics.

One of the stories included in the latter is *The Life Cycle of Software Objects*, that deals with software pets, their evolution and the relationship between these pets and their users. Without going into details for each of the books in Fig. 1, we suggest that you have a look at them because their stories deal also this type of topics.

As mentioned earlier, MIT press is becoming very interested in publishing science fiction narrative that helps teach courses on techno-ethics, because there is a shortage of that type of materials. For instance, not long ago they published *Robot Futures* by Illah Nourbakhsh, who is a professor in computer science at Carnegie-Mellon University, and he teaches a course based on these short stories. Along this line, MIT Press offered to translate my novel, originally written in Catalan and entitled *La Mutació Sentimental* (whose literal English translation would be ”The Sentimental Mutation”) and published it with the title *The Vestigial Heart*, together with ancillary materials to teach a course on ”Ethics in social robotics and AI.

This course and materials will be described in the second part of the tutorial, but before let us open for questions, comments and a bit of debate.

7 Question answering - first round

Question from audience: *I have the feeling that sometimes series like 'Black Mirror', for example, instead of making people think about some issues, in some way they normalize some dystopian future, and I don't know if you have some thoughts on this or some hints.*

Carme's answer: This is a very nice and relevant question, since it is true that most science fiction narratives are dystopias, not precisely utopias. But I would say that authors who write these dystopias do it because they are optimistic. Why are they optimistic? Because they raise these issues to make people aware of what may happen, and they are optimistic in the sense that they think people can manage to change the course of history towards more utopian futures, so that they can somehow take measures to avoid bad uses of technology and head towards good futures instead. You may say this is turning things upside down somehow, but not, I really believe this, I myself wrote this book, *The Vestigial Heart*, which you can interpret that is a dystopia.

I will say a bit about the plot because it will come out in the second part of the tutorial. It's about a 13 years old girl, from our times, who has a terminal illness and her parents decide to 'cryogenize' her, meaning to frozen her somehow. She wakes up 100 years from now and is adopted by a mother of that future, where each person has a personal assistant, that is a robot, and where several things in the minds of people have changed, have evolved, for instance some feelings, some emotions have dried out, have disappeared, and other capacities have emerged. This girl of course remembers very well her biological parents, her teachers, everyone in her past life, and it's a big contrast for her in face of the new society, with her new mother and teachers, and so on. So, you can think that this is a dystopia, and it is, because she has a bad time at the beginning in this future society until other things happen as the plot develops. But I wrote it precisely because I'm optimistic, and I think we can orient our research towards what can be more beneficial to society. I think in general the research community is oriented to try to increase the well-being of people, to remove or avoid discrimination, to try to close digital gaps, and so on, but sometimes they are not aware of the implications.

Thus, I wrote that novel, about 13 years ago, and also other novels afterwards, first to reflect myself on what type of research I would like to pursue with my group. We have group discussions regularly, once a month, about what type of projects we would like to develop. And I hope that science fiction and these dystopias can be useful in this way of picturing and making people aware that the technology they are devising will shape humanity, will shape the way we will become, and in this sense I think dystopias are written by optimists, as I mentioned. But it's very important to accompany, not just leave students or general public watch and greet these dystopias without reflecting together with them about the consequences. I really like the use science fiction narrative to raise awareness, not just as an entertainment and to give scary views of what the future may look like.

I hope to have answered your question, which is very pertinent, since it is true that this type of dystopias attracts people, readers and movie watchers more than happy utopias, which are perhaps less exciting, but this is something we have to go with, that we cannot remedy, and the media sometimes amplify these dangerous futures to get larger audiences.

Question from audience: *I just read the 'I, Robot' series and a lot of these stories that take place in those far away space stations or something that's clearly not possible today, so do you think it is more complicated to read those as an ethical exercise and then transfer them to the world as it is today? Does the fact that the setting is so far in the future make it more complicated to actually use it as a useful exercise rather than just some story in the future, especially in teaching, if you're using it as teaching material?*

Carme's answer: Well, I like better the type of science fiction that amplifies the problems we have in the current society, in the current world. Not so much these space operas that do not have much contact with our reality. I'm more in favor of social science fiction, like this *Black Mirror* series I was mentioning, which takes a technology that is available today and maybe enhances it a lot to extremes, but that amplifies essentially problems that are now embryonic, that are really in a seed nowadays. Anyway, I will later explain an experience I had recently in which several writers were contacted by a publisher to write some short stories based on Mars. Well, now there are people that go to the space for tourism and things like that, and we were writing stories based on Mars, but stories that could happen in a new settlement today and deal with the needs these people would have. For instance, I wrote about robot companions in Mars, and this is not space opera, it is another type of amplification, as I mentioned, of things that happen here on Earth, but in another society that is starting from scratch, so with new politics, new social organization, new everything. This permits imagining other worlds, another type of relations with technology. I'm in favor of this, it is not many years into the future, but just now going to the Moon and setting up a new society. I like very much these simulation strategies within groups to hypothesize futures that could happen, those that we would like to happen and others that we would like to avoid. But I agree that some of these stories in the very distant future are not really very relevant to the problems ethics today wants to address and tackle. So, we have to choose somehow.

Question from audience: *my question is about user acceptance, how do you implement that? Based on your research on assistive robotics and the way science fiction has amplified possible risks, how do you consider user acceptance in this context?*

Carme's answer: Well, this is one of the topics of the materials I have developed, acceptance of robots by people, and it is very important for the type of research we are doing. As you saw, we do a lot of human-robot interaction and sometimes with people with disabilities. For instance, this experience I told before about a robot supplying cognitive training to people with mild cognitive disabilities, that we performed during the pandemic, taught us a lot about this acceptance. We were a bit hesitant that our robots could scare these people, or that they would just refuse to interact with them because they were strange or things that they were not used to interact with. And they were elderly people, more than 70 years old with mild disability. So, it was something we thought could be difficult, and it was just the other way around. Actually, we really struggled to make those robots very friendly, they have a face with cartoon-like traits and have a repertoire of expressions of consent, engagement, happiness, and so on. We tried to make them as engaging as possible and it was a success in the sense that the patients wanted to interact more and they were not scared at all. Maybe this was favored because they had been confined for several months, not interacting, and this was an activity that attracted them a lot when they could get out and go to the Alzheimer Foundation and interact with these robots. Thus, we think that if we struggle to make these robots friendly and very respectful to people, the acceptance would be high. But, as I said, it is an important topic many studies have been touching upon, as will appear in the next section.

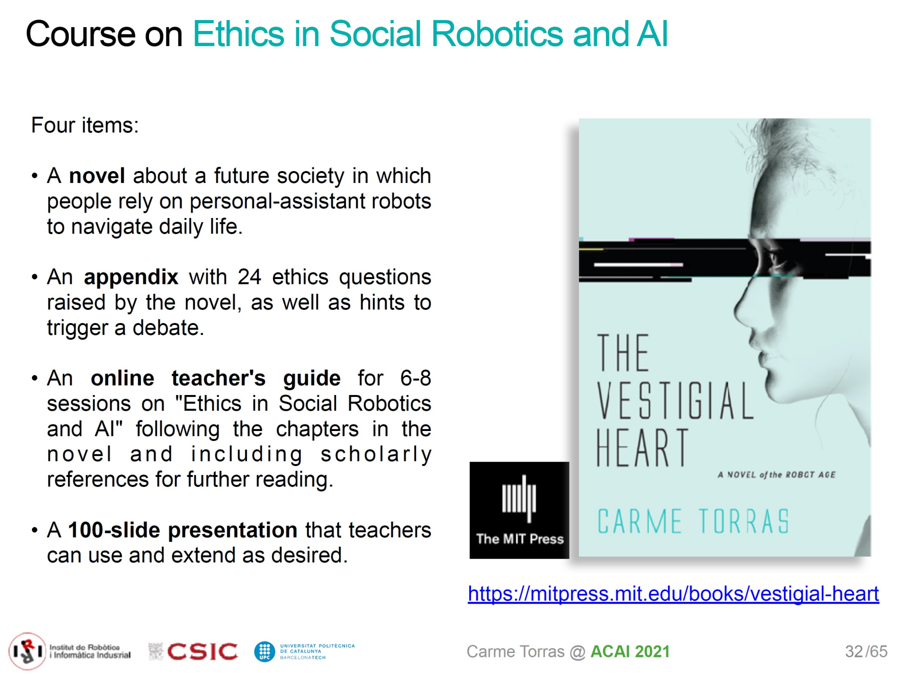
8 Plotting, writing and teaching with *The Vestigial Heart*

Now let's come to the second part of the tutorial, in which we will explain how to teach ethics using science fiction, and in particular with the book *The Vestigial Heart*, as well as some lessons we have learned and suggestions for future initiatives.

The Catalan version of the book appeared 13 years ago, then it was translated into Spanish, and finally, three years ago, it was translated into English by MIT Press, as already mentioned. The leitmotif of this novel is "it is the relationships that we have constructed which in turn shape us". This was said by the philosopher Robert Solomon in his book *The Passions*. He said so in relation to people, namely our relationships with our parents, with our friends, with our teachers, with our children, with everyone, shape the way we are. We would be different if we had interacted with other people, different persons. Then, since we have so much interaction with technology nowadays, it is very important that we think very deeply what technology we would like to develop, because it will shape the way we will be and especially future generations. Therefore, we have a tremendous responsibility in the type of technology we are developing.

So, what is the techno-ethics course based on *The Vestigial Heart* about? The book includes the novel, an appendix with 24 ethics questions raised by scenes in the novel, and some hints for a debate. These are all contained in the physical book, the paper book. And then there are online ancillary materials: a guide for teaching 8 sessions on ethics in social robotics and AI, following the chapters in the novel and including references for further reading; and there is also a hundred-slide presentation that teachers can adapt, extend and use as they wish (Fig. 2). These materials are being used in several universities in the States, in Europe and also in Australia. They can be downloaded from this address [1] free of charge, by just filling up a form with some data on the envisaged usage.

These materials have also been translated into Catalan and Spanish, as well as adapted to high-school by Pagès Editors [13].



Course on Ethics in Social Robotics and AI

Four items:

- A **novel** about a future society in which people rely on personal-assistant robots to navigate daily life.
- An **appendix** with 24 ethics questions raised by the novel, as well as hints to trigger a debate.
- An **online teacher's guide** for 6-8 sessions on "Ethics in Social Robotics and AI" following the chapters in the novel and including scholarly references for further reading.
- A **100-slide presentation** that teachers can use and extend as desired.

https://mitpress.mit.edu/books/vestigial-heart

Carme Torras @ ACAI 2021 32/65

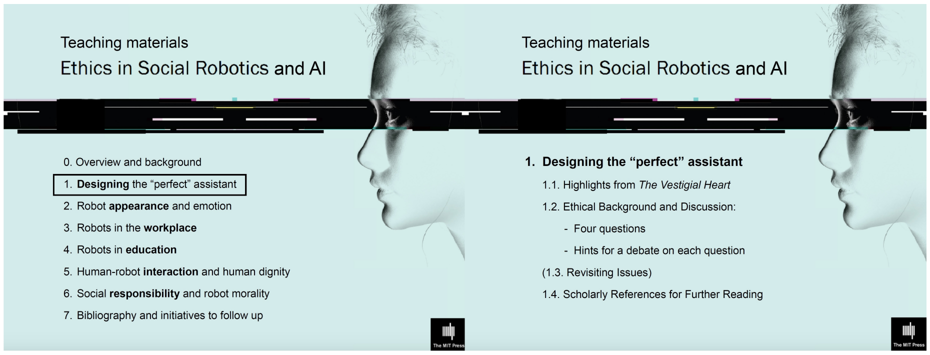
THE VESTIGIAL HEART
A NOVEL of the ROBOT AGE
CARME TORRAS
The MIT Press
Institut de Robòtica i Informàtica Industrial
CSIC
Universitat Politècnica de Catalunya BarcelonaTech

Fig. 2. Techno-ethics course materials.

There are six main chapters in these materials: Designing the 'perfect' assistant; Robot appearance and emotion; Robots in the workplace; Robots in education; Human-robot interaction and human dignity; and Social responsibility and robot morality; and they are wrapped up with an introductory and a concluding chapter.

Starting with the introductory one, first there is a quick overview of several ethical theories from the history of human ethics that can be applied to robotics. There are many quite diverse theories, like utilitarianism, deontologism, virtue ethics, social justice, common goods, religious ethics, information ethics, and many of them are somehow contradictory. Since no single theory is appropriate for addressing all the issues (specially the new ones), along the materials, a pragmatic option is taken, which has been termed 'hybrid ethics' and is advocated by Wallach and Allen in their book [14]. It entails combining top-down theories, those applying rational principles to derive norms, like virtue ethics for instance, with bottom-up theories, which infer general guidelines from a specific situation. This pragmatic option is very common nowadays and, along this line, since there are different cultures, countries and values for each of them, the general theories need to be applied and instantiated for different cultures, values and circumstances.

We will give a glimpse of the chapters and the issues addressed in each of them. More details can be found in the tutorial slides [15]. All chapters have the same structure (Fig. 3). They start with some highlights from the novel, then some ethical background is provided and 4 questions are posed to students. Next, hints for a discussion are provided mainly for the teacher to foster a debate on each question. In later chapters, there are revisited issues that reappear from different viewpoints, and finally many scholarly references for further reading, and for suggesting projects to students are provided.



Teaching materials
Ethics in Social Robotics and AI

0. Overview and background
1. **Designing** the “perfect” assistant
2. Robot **appearance** and emotion
3. Robots in the **workplace**
4. Robots in **education**
5. Human-robot **interaction** and human dignity
6. Social **responsibility** and robot morality
7. Bibliography and initiatives to follow up

Teaching materials
Ethics in Social Robotics and AI

1. Designing the “perfect” assistant

1.1. Highlights from *The Vestigial Heart*
1.2. Ethical Background and Discussion:
- Four questions
- Hints for a debate on each question

(1.3. Revisiting Issues)
1.4. Scholarly References for Further Reading

Fig. 3. Chapters and structure.

Before entering the six main chapters, let us summarize the plot of the novel. Celia, a-thirteen-year-old girl cryogenically frozen because of her terminal illness, is cured and brought back to life 100 years later 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.

8.1 Designing the ’perfect’ assistant

Doctor Craft at the beginning of the book is very obsessed in getting the prosthesis because he is getting old, he was very creative in the past, and now he wants his creativity to be stimulated. He makes this reflection: “... A good choice of stimuli, that’s the secret to wellbeing. [. . .] we can’t change man or turn his brain upside down, we can’t modify even the smallest reaction. The only option is to control his surroundings, control what he feels through the stimuli he receives. A key idea, but when he presented it as the leitmotif of the new line of robots, no one gave it a bit's worth of notice". Note that he is the owner of the company and the main chief that provides ideas. "Too simple, they said. How short-sighted! [. . .] they couldn't tailor-make a ROB for everyone; they had to come up with a generic ROB that was highly adaptable and, most important of all, one that could achieve a very fast adaptation. If it took one week for a ROB to work out how to wake up its PROP", its PROP is its user, "or how much sugar to put in his coffee, the whole idea would go down the drain". Thus, he is advocating for having very personalized robots that adapt to the user immediately. And, as mentioned, he is obsessed with having a creativity prosthesis for him. And the quotation continues: "what he [Doctor Craft] wants is a creativity prosthesis. Or an assistant [. . .]; something that would stimulate him to think differently, that would warn him when he started down well-worn paths and would show him the promising forks in the road, those susceptible to innovation".

These quotations try to make the distinction: what he wants for himself is something that stimulates him, while what he wants to sell are robots that adapt to people, and give them what they want, without having to think or do much. This exemplifies a double moral, which we discuss and leads to a debate on what is a perfect assistant, giving rise to several questions for debate. The four questions posed in this chapter are:

1. Should public trust and confidence in robots be enforced? If so, how? This is one of the questions asked and discussed in the first round of question answering above.

2. Is it admissible that robots/applications be designed to generate addiction? Because you know that apps are sometimes designed to generate addiction, especially from young people.

3. Should the possibility of deception be actively excluded at design time?

4. Could robots be used to control people?

When giving a course, these questions are raised and then, as already said, in the teacher's guide there are some hints for a debate, but here we will continue to other chapters. We cannot go over the eight sessions in just one hour!

8.2 Robot appearance and emotion

In the second chapter, which is about robot appearance and simulating emotions, some highlights are as follows: "as a birthday present, Lu [who is the adoptive mother] gave me [Celia] a robot. [. . .] it has a kind of head with no nose, mouth or ears, it just has two cameras, and a screen embedded in its chest. It's called ROBbie." In another scene: "Celia, touched by the words, looks for his eyes: no friend had ever sworn their loyalty so convincingly, but two black holes bring her back down to earth. Though not entirely, [. . .] she watches the robot out of the corner of her eye and it pleases her to see his dignified posture, gently swinging his strong, shiny arms. It feels good to walk along beside him, she feels protected, she can trust him. What does it matter that he doesn't have eyes, people don't look at each other anymore anyway." This is about trust generation by robots, and also the contrast with the society of the times, 100 years from now, in which people don't look at each other, they are so little empathetic that they don't really relate. This topic usually provokes strong debates.

The questions triggered by these and other highlights (Fig. 4) are:

1. How does robot appearance influence public acceptance?

2. What are the advantages and dangers of robot simulating emotions? This is another big topic: robots simulating emotions and people getting cheated or getting deceived because of this; ethics regulations and education should try to avoid this confusion in grown-ups and, especially, in children.

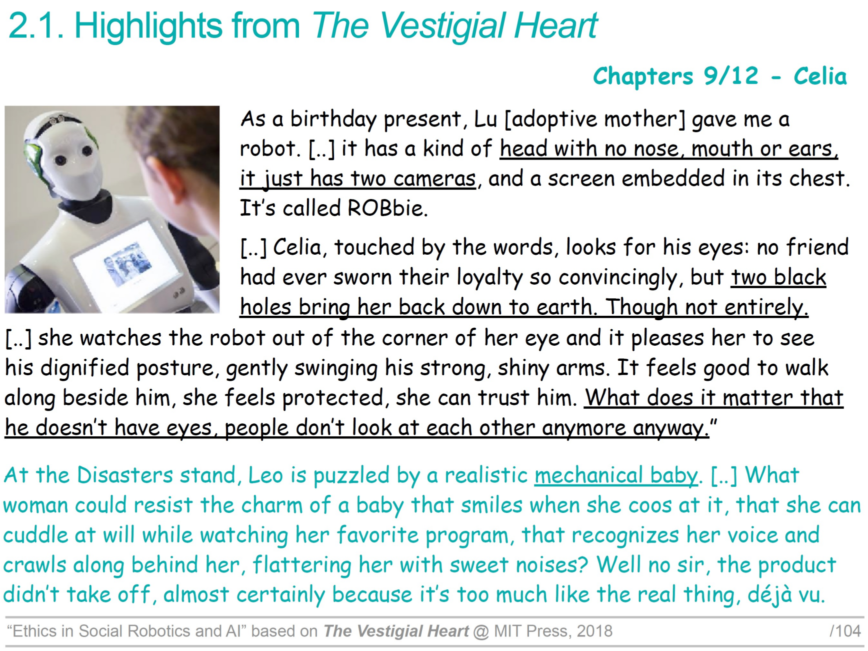
3. Have you heard of/experienced the "uncanny valley" effect? This effect is that people get more attached to robots, or to artificial creatures, the more anthropomorphic they are. For instance, the robot in the movie Wall-E has a kind of head with eyes and two things similar to arms, wheels that resemble legs, so it has a kind of anthropomorphic cartoon-like appearance. As robots get more and more anthropomorphic, both physically and also cognitively, people get more attached to them because we empathize more somehow. The empathy and the attachment increases up to a point in which an extreme similarity to people, but with an strange look, provokes repulsion. The growing acceptance curve suddenly goes down abruptly, this is the uncanny valley effect. This has to be taken into account by robot designers, and it is the reason why robots often have a cartoon-like face, because this engages people without cheating.

4. Should emotional attachment to robots be encouraged? Well, this depends on the situation, so it is discussed at length in the materials.

8.3 Robots in the workplace

Here we will just highlight one passage from the novel. In the materials, more quotations are included for each chapter, and everything is well specified for instructors to be comfortable and ease their teaching. The high says "Leo looks up at the ever-watching electronic eyes [in his workplace] and thinks how lucky he is that they can't read his mind." This is anticipating that if, in the future, we wear implants, our boss may be able to read what we are thinking. And this poses questions of privacy, and also of property rights. "ROBco [which is Leo's robot] is waiting expectantly next to him, ready to act as his assistant." And Leo says: "I can't turn off and back on again and pick up where I left off, like you do [because robots are very quick at changing the context], see if you can finally build that into your model". They are cooperating at work and, of course, robots have to have a model of their user in order to be useful, so this raises many questions also of human-robot cooperation and these are discussed.

A theme to debate is how to define the boundaries between human and robot labor in a shared task, so that not only throughput is maximized, which is what the owner of the company wants, but, more importantly, the rights and dignity of professionals are preserved. This relates to some passages in which Leo struggles on two fronts: his privacy and intellectual property rights may be violated by the device they have that can somehow read his mind; and he struggles also to make his robot "understand" that they have different skills and, to optimize collaboration, they need to do what each does best and communicate on common



2.1. Highlights from *The Vestigial Heart*

Chapters 9/12 - Celia

As a birthday present, Lu [adoptive mother] gave me a robot. [..] it has a kind of head with no nose, mouth or ears, it just has two cameras, and a screen embedded in its chest. It's called ROBbie.

[..] Celia, touched by the words, looks for his eyes: no friend had ever sworn their loyalty so convincingly, but two black holes bring her back down to earth. Though not entirely. [..] she watches the robot out of the corner of her eye and it pleases her to see his dignified posture, gently swinging his strong, shiny arms. It feels good to walk along beside him, she feels protected, she can trust him. What does it matter that he doesn't have eyes, people don't look at each other anymore anyway."

At the Disasters stand, Leo is puzzled by a realistic mechanical baby. [..] What woman could resist the charm of a baby that smiles when she coos at it, that she can cuddle at will while watching her favorite program, that recognizes her voice and crawls along behind her, flattering her with sweet noises? Well no sir, the product didn't take off, almost certainly because it's too much like the real thing, déjà vu.

"Ethics in Social Robotics and AI" based on ***The Vestigial Heart*** @ MIT Press, 2018 /104

Fig. 4. Highlights to discuss about robot appearance and emotion.

grounds. This is trying to get to the utopia that robots and people collaborate using the best of themselves. The questions for this chapter are as follows.

1. Would robots primarily create or destroy jobs? This is the million-Euro question, that since the industrial revolution the media raises constantly, now concerning robots and AI, so this is discussed at length in the materials.

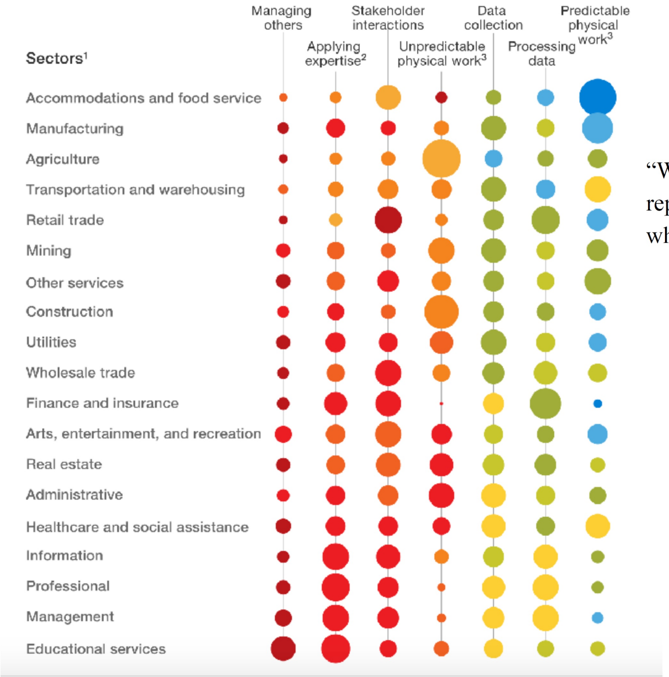
2. How should work be organized to optimize human-robot collaboration?

3. Do experiments on human-robot interaction require specific oversight? This is so, there are specific rules and guidelines to perform experiments with humans, in particular in EU projects. All these ethics issues need to be taken into account and, for example, consent forms and information sheets must be supplied to participants in experiments.

4. Do intellectual property laws need to be adapted to human-robot collaborations? Of course, yes, so these are discussed.

Just a glimpse on some data provided in the materials. In a wide survey performed by Frey and Osborne already in 2013 [16], about 47% of the total US employment was then at risk of being automated. And high wages and educational attainment have negative correlation with such risk, since the tasks that are being mechanized or automated are usually those that have low quality value. Three features that preserve jobs for humans are: (i) requiring negotiation, since humans are very good at this and the required capacities are difficult to be automated; (ii) needing creative solutions in front of unknown problems, unexpected situations, for instance doing maintenance of facilities that have broken down; and (iii) entailing empathizing with the situation of other people. Thus, jobs related to community and social services have the lowest percentage chances of being automated, whereas telemarketing has the highest percentage, this is the conclusion of that survey.

In another study, McKinsey & Company [17] plotted several jobs and, after interviewing many people, came up with an ordered list of jobs from the easiest to the hardest to be automated (Fig. 5). Among the hardest is education because teachers need to empathize with the students, motivate them, and transmit the life experience they have and their passion for knowledge.



"Where machines could replace humans—and where they can't (yet)"

McKinsey & Company

2017

Fig. 5. Adapted from [17].

It is worth pointing out that what is susceptible of being automated are tasks, rather than entire jobs. A job description consists of several tasks, those most repetitive and predictable can be automated easily, but jobs often also entail other tasks that require intuition, creativity and empathy, the three human traits that were highlighted as the most difficult to be automated. This is well aligned with what happens in the novel, thus providing good examples and favoring the discussion of these issues.

Another conclusion from a related study is that “the best solutions are always going to come from minds and machines working together”, thus taking the best of both type of agents. And the challenge is, of course, not to fall into complete technological dependency. There is a big risk of being so dependent on machines that if there were a blowout we could not recover, so it’s important to retain human autonomy and avoid this dependency.

8.4 Robots in education

Celia goes to a school in which teachers do not know anything, they just search and help students to search because the net, called EDUsys, which is more sophisticated than ours, has all the knowledge, so that students do not have to learn to think, they just learn to search. In a passage from the novel, we are told that the teacher has labeled Celia a rebel because, ignoring his advice, she insists on competing with machines. Why? Because apparently she has not understood the net’s search mechanisms and, faced with a question, she stops and thinks about it, trying to make up an answer. Therefore, Celia gets very bad grades and is questioned all the time, just because, when faced with a problem, she tries to solve it by herself, not through search mechanisms. Since she does badly at school, her mother puts her a home teacher, Silvana, who is another big character in the novel. Silvana will have to work hand in hand with ROBbie, Celia’s robot, because everyone knows that if the child turns out to be a rule-breaker, like Celia, the robot must learn to restrain them, to counteract their impulses to break rules. Robots are customizable for a reason, they have to complement their PROPs to make a good team user-robot, but that is the last thing Silvana expected to hear, that she will have to train a robot, because she is against technology.

In the novel there are the “pro-techno’s”, who are pro-technology, like Leo, Doctor Craft and all the people that work devising robots and technology in general, and the ”anti-techno’s”, like Silvana, who form a community called ”the ComU”, in which they refuse to be assisted by robots, they want to be just fully autonomous humans. The confrontation between them permits exemplifying many of the ethical issues that are raised by digital technologies nowadays. The questions raised in this chapter are as follows:

1. Are there limits to what a robot can teach?
2. Where is the boundary between helping and creating dependency?
3. Who should define the values robot teachers would transmit and encourage?
4. What should the relationship be between robot teachers and human teachers?

8.5 Human-robot interaction and human dignity

The confrontation between pro and anti-technology perspectives can be observed in some dialogues between Silvana and Leo, who often get into this fight. In a passage from the novel Silvana says: ”machines that augment human capabilities seem like a great idea to me: without remote manipulators surgeons couldn’t operate on a microscopic scale and, without INFerrers, we’d take too long overthinking the consequences of our decisions . . . it’s ROBs that I reject, and the personal link that is established between them and their PROPs [their users] that ends up hogging people’s most intimate time and space. You said it yourself [Leo]: you don’t need anything else . . . and, in the end, you become wooden like them.”

In another passage Leo is discussing with her and he says: “I don’t understand [what you say]. All ROBs are developed to serve people.”, and Silvana answers: “Exactly. It’s just that the service is often poisoned. Why do you think we are against those mechanical contraptions? [...] Because we’re snobs? Well, no. [...] Overprotective robots produce spoiled people, slaves produced despots, and entertainers brainwash their own PROPs. And worst of all you people don’t care what happens to the rest of us as long as they [the robots] sell”.

The questions for this chapter are as follows:

1. Could robot decision-making undermine human freedom and dignity? Of course, this is a big risk and we, who are working in this field of human-robot interaction, especially in the assistive context, are very aware of it and discuss it often.

2. Is it acceptable for robots to behave as emotional surrogates? If so, in what cases? It could be cases in which faking emotions can be helpful, for instance, as we said, to engage people, as long as people are aware that they are interacting just with a machine. Machines showing happiness or sadness in particular circumstances can facilitate a friendly interaction. In some cases, like in emergency situations, a robot showing emotions can be more convincing to warn or guide people than an expressionless robot. But these are very specific circumstances, in general faking emotions, we think, should be avoided or used very cautiously.

3. Could robots be used as therapists for the mentally disabled? It depends on the circumstances and, of course, it has to be under human therapist control.

4. How adaptive/tunable should robots be? Are there limits to human enhancement by technological means? This will be briefly addressed in the Conclusion section.

8.6 Social responsibility and robot morality

This is the subject of the last chapters in the novel, but we did not include any highlights from them to avoid making spoilers in case some readers want to read the novel with some intrigue. The questions for this chapter are as follows:

1. Can reliability/safety be guaranteed? This is a main issue. How can hacking/vandalism be prevented? Think of autonomous vehicles, for example.

2. Who is responsible for the actions of robots, of vehicles, of programs, of everything? Should moral behavior be modifiable in robots, machines, programs?

3. When should a society’s well-being prevail over the privacy of personal data?

4. What digital divides may robotics, artificial intelligence, etc., cause?

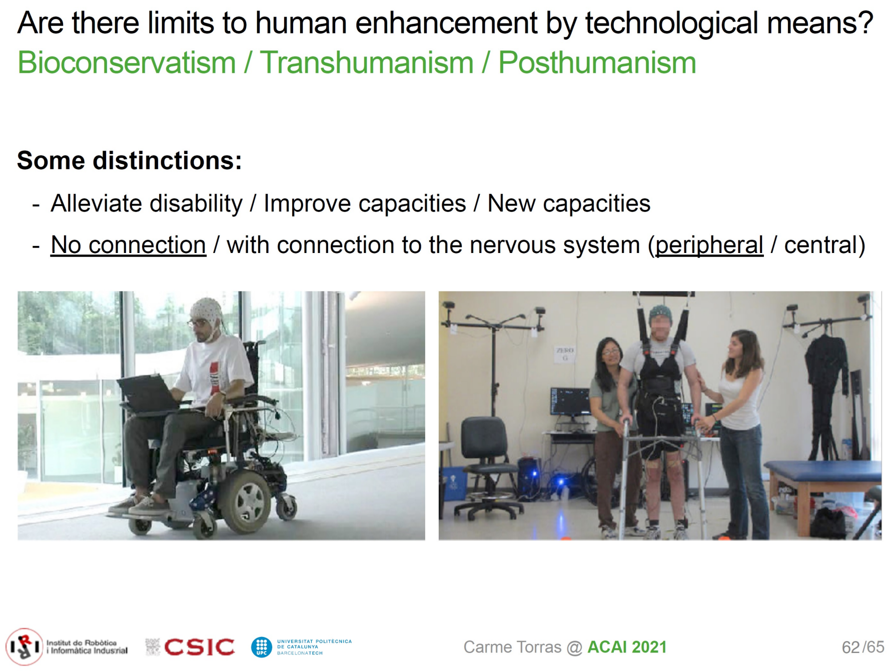
9 Concluding remarks and other writing initiatives

We would like to end this chapter by proposing and showing two experiences in collective writing. One of them was already mentioned, and we will explain another one that is related to a question we posed in section 8.5: are there limits to human enhancement by technological means? Actually, this is a question in posthumanism which now is raising a lot of discussion in the media, and there is a debate between bioconservatists, transhumanists and posthumanists on whether this human enhancement is ethical, in what circumstances, and what social implications it would have.

Thus, there was an experiment during the pandemic in which nine writers, Carme among them, were asked to picture the Barcelona of 2059. In particular, we did a lot of worldbuilding by virtually getting together and trying to imagine what would happen in Barcelona in that year, and we pictured a dystopian Barcelona city that had degenerated a lot in all fronts (sustainability, poverty, discriminations). But there was an island in the seafront, on a platform, that we called Neo Icària, in which there was a utopian society whose members had to sign a contract with a company that ruled the island. The contract guaranteed them free heatlhcare, jobs that they liked and everything else they needed, but they had to give up their privacy, accepting to take an implant (chosen among three options) yearly, which was used for the company to monitor their behavior all the time, physically and psychically. Some people signed, whereas others chose not to. And some of the problems this entails are explored in the book, which has been very successful and has been generating many discussions. It was a very nice experience of writing together with other authors and getting into this discussion on bioconservatism, transhumanism, and posthumanism (Fig. 6).

One of the issues that were raised in the plot is related to people with disabilities, who wanted to enter the island even if they were monitored. But implants can not only palliate disabilities, but also provide new capacities, like more strength, exoskeletons, night vision, etc., which could, for instance, be used for military purposes. Another distinction worth making is whether this technology is invasive, i.e. connected to the nervous system, or not. But even when it is not connected to the nervous system, it has implications in our brain, because our brain is plastic. For instance, there was a real experiment of putting a sixth finger in a hand, which could be handy for playing guitar or for doing some works at home. The extra finger was activated by the muscles in the forearm but not connected to the nervous system. In the motor-cortex we have a mirror of our five fingers, but the brain is so plastic that after continuous usage of that sixth finger, it was also mirrored in the motor-cortex. This tells us that we should be very careful of what technology we are using, because it is really shaping not just our way of behaving, but also our brain and nervous system.

Another collective experience was the one we mentioned earlier of imagining that now it was possible to travel to Mars and establish a new society. This was open to utopia, to make a society that was more egalitarian, guaranteeing well-being to everyone, without discrimination, without any of the issues we handle here on Earth nowadays. We got together for worldbuilding and it was a very



Are there limits to human enhancement by technological means?
Bioconservatism / Transhumanism / Posthumanism

Some distinctions:

- Alleviate disability / Improve capacities / New capacities
- No connection / with connection to the nervous system (peripheral / central)

Institut de Robòtica i Informàtica Industrial CSIC Universitat Politècnica de Catalunya BarcelonaTech Carme Torras @ ACAI 2021 62/65

Fig. 6. Some distinctions on technology-based human enhancement.

nice experience again. While some authors wrote about politics in Mars, the health system, etc., Carme wrote a short story about a moral robot, which has some embedded values, is adaptable to the user, reconfigurable, and particularly handy in Mars, with all its peculiarities. Actually, the story is very optimistic and robots are very helpful to the first inhabitants of Mars, which concludes this chapter in a positive way by suggesting that, although technology poses many challenges, it can also be very helpful and lead us to a better, more egalitarian future. So, the question we leave open is "what role should the human being and the machine play in this *pas de deux* in which we are irrevocably engaged?"

9.1 Further reading

As supplementary materials and readings, some suggestions follow. On the topic of the role of science fiction, we recommend Stephenson's *Innovation Starvation* [3] and Torras' *Science fiction: A mirror for the future of humankind* [18]. On ethics education initiatives in the digital age, we recommend the aforementioned articles [12] and [5] as well as a lesson plan for robot ethics from the website Teach with Movies [19] based on the film *Robot and Frank*. And finally, as already indicated, the novel *The Vestigial Heart* and its associated teaching materials can be found on the MIT Press website [1].

10 Question answering - second round

Question from audience: *hi, thank you for your talk. I was wondering if you think there is an issue with the fact that most people actually get their information from social media, from newspapers, from these books, but not directly from actual science articles about the development of robots or AI that is currently going on. Especially considering that we have a bit of an iffy time with people's trust in science, more so lately with the pandemic when there has also been a misinformation time going on at the same time. So, just as a scientist, what is our responsibility in this case? How do we make sure that the information that people actually get is consistent with what is going on in science? And also, because of this disconnection between science and science fiction, how do we make sure that science fiction doesn't villainize science in a way that we are somehow endorsing these dystopic futures that are presented in novels? Even if they are only a science personal inspiration or most of the time, as you mentioned, they are rather what not to do than what to do. Yeah, this was quite a lengthy question, but I hope I was clear in my concerns about it.*

Carme's answer: yes, you were very clear and these are very important concerns, especially for us working on AI, robotics, and this type of human-machine interaction that is growing fast, especially because we interact with mobiles every day. I have many answers to your questions. One is that we have a responsibility, as scientists, to disseminate correct views of what are the benefits and risks of technology. Because sometimes there is a lot of hype going on saying that technology can do a lot more than it can, both in the positive way and in the negative way. For instance, the idea that machines will get conscious, more intelligent than humans, so they will take over and humans will be extinguished. I don't think machines will get conscious, at least not in the next 50 years, as some are foreseeing. Thus, we should be very rigorous in the information we supply to the media not to hype in any of the two directions. Another answer is that we should promote techno-ethics education at all levels. I was advocating especially for the designers of this technology: engineers, computer scientists, university students in technological degrees, and so on. But also I would like to enforce such education in high school, as I did, and in primary school, because from the very beginning children need to know about how to use technology for their own benefit, not for the benefit of companies and not to get addicted and fall in the traps that these big corporations are setting, which make us click constantly and be dependent on technology.

Therefore, I am enforcing in my local context that primary schools have also this type of courses and education in ethics in technology. But sometimes the main barrier are teachers that are not very familiar with technology, and they themselves are somehow afraid that children will know more about apps than them. There are some experiences in which children teach teachers how to manage these apps, how to deal with technology, because they know better. And teachers give them clear criteria on how to use this technology for their benefit and advise them about the risks also. In this way, there is a mutual interchange of information and values, which I think is very positive. And the same for the general public, I am constantly accepting to participate in all types of forums and outreach activities to disseminate this need for ethics and to accompany all this science fiction narrative and all this science and technology dissemination with these values and these ethics guidelines.

Finally, I am also enforcing a lot the confluence of technology and humanities, this being why I am organizing forums on this confluence at conferences and also in all types of contexts. My university, the Technical University of Catalonia, only has technology degrees, so there are no humanities, no medical degrees being taught there. Thus, for students to take humanities subjects, an agreement has been made with another university that teaches humanities, so that students in engineering can take credits in humanities for their degrees. In sum, I think it is important to enforce this confluence at all education levels.

References

1. Torras, C.: The Vestigial Heart, together with instructor resources, `https://mitpress.mit.edu/books/vestigial-heart`. Last accessed 12 Jan 2022
2. Many Worlds. Nature **448**(7149), 1–104 (2007)
3. Stephenson, N.: Innovation starvation. World Policy Journal **28**(3), 11-16 (2011)
4. Center for Science and the Imagination, `https://csi.asu.edu/`. Last accessed 12 Jan 2022
5. Torras, C.: Assistive robotics: Research challenges and ethics education initiatives. DILEMATA: International Journal of Applied Ethics **30**, 63–77 (2019)
6. Web 2.0 Suicide Machine, `http://suicidemachine.org/`. Last accessed 12 Jan 2022
7. ETER9, `https://www.eter9.com/`. Last accessed 12 Jan 2022
8. HereAfter, `https://www.hereafter.ai/`. Last accessed 12 Jan 2022
9. Replika, `https://replika.ai/`. Last accessed 12 Jan 2022
10. Legathum, `https://legathum.com/`. Last accessed 12 Jan 2022
11. Grosz, B.J., Grant, D.G., Vredenburgh, K., Behrends, J., Hu, L., Simmons, A., Waldo, J.: Embedded EthiCS: integrating ethics across CS education. Communications of the ACM **62**(8), 54–61 (2019)
12. Burton, E., Goldsmith, J., Mattei, N.: How to teach computer ethics through science fiction. Communications of the ACM **61**(8), 54–64 (2018)
13. Torras, C.: La Mutació Sentimental, `https://www.pageseditors.cat/ca/guia-didactica-la-mutacio-sentimental.html`. Last accessed 12 Jan 2022
14. Wallach, W., Allen, C.: Moral machines: Teaching robots right from wrong. Oxford University Press, Oxford (2008)
15. Torras, C.: Writing Science Fiction ACAI Tutorial, `https://www.iri.upc.edu/people/torras/2021-10-14\_ACAI-Tutorial\_CarmeTorras.pdf`. Last accessed 12 Jan 2022
16. Frey, C.B., Osborne, M.: The future of employment (2013)
17. McKinsey & Company: Jobs lost, jobs gained: What the future of work will mean for jobs, skills, and wages, `https://www.mckinsey.com/featured-insights/future-of-work/jobs-lost-jobs-gained-what-the-future-of-work-will-mean-for-jobs-skills-and-wages`. Last accessed 12 Jan 2022
18. Torras, C.: Science-fiction: A mirror for the future of humankind. IDEES **48**, 1–11 (2020)
19. Teach with Movies: Robot and Frank, `https://teachwithmovies.org/robot-and-frank/`. Last accessed 14 Jan 2022