Robotics and AI meet the Humanities:
Some initiatives for ethics education and dissemination

Carme Torras[*]
Institut de Robòtica i Informàtica Industrial (CSIC-UPC)
c/ Llorens i Artigas 4-6, 08028-Barcelona, Spain
http://www.iri.upc.edu/people/torras

1. Introduction

The influence of the humanities on the study of technological subjects —such as robotics, biomedical engineering, artificial intelligence, data science or biotechnology, to name a few— needs to rapidly grow, for the simple reason that these technologies are becoming a part of humanity: assisting, interacting, and enabling people in an increasing number of ways in daily life.

Although computer and robotic technologies are often considered a further step in a social transformation that started with the agricultural and industrial revolutions, they introduce a qualitative difference. It is no longer only a matter of mechanizing heavy and repetitive tasks in farms and industries, or electrical appliances freeing up people's time for use in more creative and enjoyable ways. The difference lies in the fact that these new technologies enter domains so far considered exclusively human, such as decision-making, emotions, and social relationships, which may compromise human values, decisively shape society and our way of life, and ultimately influence the evolution of humankind.

Thus, the trend towards more and more specialization that has dominated higher education in the last decades needs to be somehow counteracted by

[*] Carme Torras (www.iri.upc.edu/people/torras) is Research Professor at the Robotics Institute (CSIC-UPC) in Barcelona, where she leads a research group on assistive robotics. She is IEEE and EurAI Fellow, and member of Academia Europaea. Convinced that science fiction can help promote ethics in digital technologies, her novel *The Vestigial Heart* (MIT Press, 2018) has been published together with online materials to teach a course on "Ethics in Social Robotics and AI".

adopting a wider view that takes into consideration the social implications of the technologies being studied. As an example in this direction, 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 (ACM/IEEE CS curricula 2013), one of which is "Social Issues and Professional Practice", so that "students develop an understanding of the relevant social, ethical, legal and professional issues". The need to incorporate the study of these non-technical issues into the ACM curriculum was formally recognized in 1991.

Researchers and professionals in these technological areas are also becoming aware of the importance of joining efforts with social scientists, lawyers, philosophers and anthropologists, among other humanities scholars, and interdisciplinary teams are becoming more and more common practice as well as much appreciated. Therefore, having a cross-disciplinary engineering education, which permits bridging language gaps with professionals from other fields and different backgrounds, will increasingly be a key feature to develop a successful career. Using theoretical frameworks from education and psychology to ground the results of an experimental study with 24 individuals, Borrego and Newswander (2008) provide some recommendations for funding agencies, professional societies, university administrators and faculty to promote cross-disciplinary engineering education.

2. Technological degrees open up to the humanities

There are many options to integrate contents from the humanities in technological university degrees, ranging from including in the syllabus a course on good professional practice, to allowing students to take some credits or a minor in a Humanities Department, to even offering a mixed degree, like the Computer Science and Philosophy degree recently launched at the University of Oxford (http://www.ox.ac.uk/admissions/undergraduate/courses-listing/computer-science-and-philosophy), which focuses on common interests in artificial intelligence, logic, robotics, virtual reality, and ethics.

It is unsurprising and most desirable that, among humanities subjects, ethics gets a prominent place in the syllabus, as it explains how human values relate to life styles, and digital technologies are already having a strong impact on our daily lives and, thus, are shaping human most important values.

Several universities in Europe, and many in USA and Canada, offer undergraduate courses related to Ethics in Technology, the Digital Age, Society and Technology, etc., which, following the above-mentioned ACM/IEEE Computer Science curricula, address issues like privacy, intellectual property, safety, reliability, autonomous and pervasive technologies, vulnerable groups, and professional ethics. Such courses are not only offered in Computer Science and Engineering degrees, but they are often taught in Philosophy departments as well.

A typical situation is that at the university of West England at Bristol, where the syllabus of the Robotics BEng degree includes a compulsory course on Ethics of Technology taught in the Philosophy department.

Among the universities regularly offering courses related to technology ethics are those of Bristol, Leeds, Sheffield, Oxford, Twente, Genoa, Pisa, Barcelona Tech, Stanford, Carnegie-Mellon, Yale, Southern California, California Santa Barbara, California Polytechnic, Georgia Tech, Texas Tech, Notre Dame, Miami, Southern NewHampshire, Lower Massachusetts, Worcester Polytechnic, Illinois, British Columbia, Carleton, Ottawa and Dalhousie, to name a few.

Information on similar initiatives in other continents is scarcer. Seoul National University offers ethics courses tailored to almost every major, as well as core courses on “Information Society and Cyberethics” and “Engineering Ethics and Leadership”. Japanese and Australian universities have similar programs. The Japanese ministry of education and science is funding a project on “the construction of robot ethics according to the practical interest of roboticists”, which should establish education guidelines on robot ethics. As another singular example, Queensland University of Technology in Brisbane offers a Robotics online course (the popular MOOCs), where one of the 12 topics covered is “Robots in society and associated ethical issues”.

In short, there is a clear need to include focused and practical ethics courses in Technological curricula. Computer Science degrees have played a pioneering role in this regard, and several universities already offer ethics courses in such degrees. Given the fast expansion of digital technologies, AI, bioengineering, social robotics, and similar subjects, it is envisaged that technology ethics courses will soon proliferate not only in Computer Science, Electrical Engineering and Philosophy curricula, but will percolate as well to related disciplines in the social and natural sciences.

3. An experience: Robotics meets de humanities

The robotics research community is well aware of this needed confluence with the humanities and many joint initiatives have been undertaken, such as the launching of research projects (euRobotics 2012; RoboLaw 2014), the publication of special issues in scientific journals (Veruggio *et al.* 2011), and the organization of open forums on “Robotics meets the Humanities” at the main robotics conferences (ICRA Forum 2013; IROS Forum 2018).

These forums are often open to the general public and gather as invited speakers not only university professors from technical and humanities departments, but also filmmakers, science-fiction writers, dancers, stage performers, and representatives from Governmental institutions, all having in common their belief that robotics development must take into account our understanding of humanity in its multiple facets.

The view I presented at those forums of how bridges between robotics and the humanities may be tended is pictured in Figure 1. Both disciplines share a modeling level aimed at discovering facts and relationships, and scholarly work of joint teams is starting to be carried out relating physical models developed by roboticists to psychological and social models studied in the social sciences. Moreover, as exemplified in the forums mentioned above, bridges can also be established at an artistic level by joining the creative sides of both robotics and the humanities, namely technological innovation and science fiction, respectively.

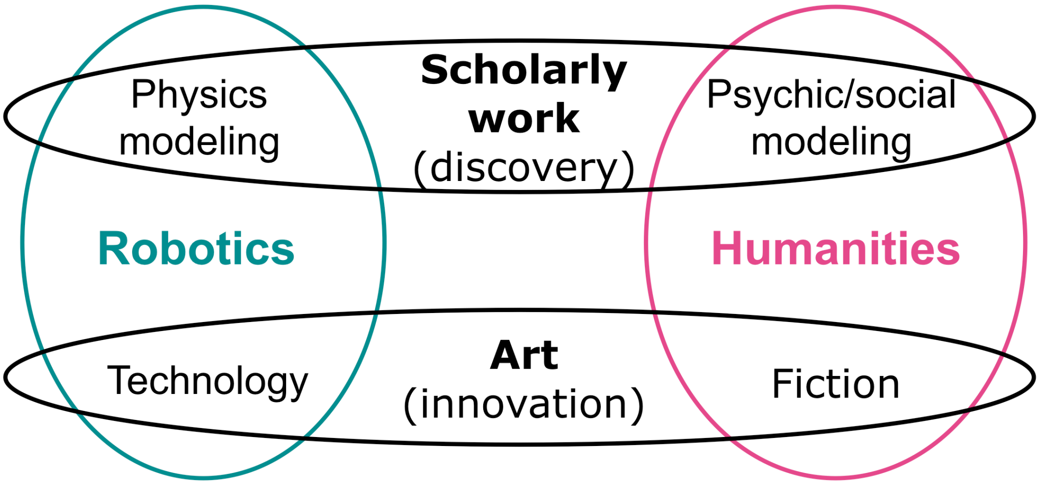


Figure 1. Graphical representation of possible bridges between Robotics and the Humanities.

As an example, in 2012 the Center for Science and the Imagination was set up at Arizona State University (http://csi.asu.edu/). The idea came from a thought-provoking speech by Neal Stephenson (2011), 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 science-fiction 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 several research groups that bring together researchers in science and humanities to devise and endeavor to achieve 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 science-fiction 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 (Rushkoff et al. 2012), 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.

The confluence of robotics with the humanities has even resulted in 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 2005). The discipline involves two main areas: legal regulation and ethical education. Regarding the former, several institutions and professional associations are developing regulations for robot designers, programmers, and users. The European Parliament (2017) released some guidelines under the general title of *Civil Law Rules on Robotics*. Other examples are those put forth by the British Standards Institution (2016) and the IEEE Standards Association (2019).

Concerning education, although Roboethics is touched upon in most courses dealing with ethics in technology and the digital society, I am only aware of a few ethics courses devoted entirely to robots. I would highlight those in the Philosophy departments of Carleton and Notre Dame universities, and those in the Computer Science departments of Carnegie-Mellon (available online) and Illinois-Springfield (an online course as well, fulfilling the Engaged Citizenship Common Experience requirement). In Europe, some such courses are also offered at Delft University of Technology, University of West England at Bristol, and Oslo University, among others.

Given the fast expansion of social robotics and intelligent systems, it is envisaged that Roboethics courses will soon proliferate in Computer Science, Engineering and Philosophy curricula, and then percolate to related disciplines in the social and natural sciences.

4. Ethics courses for technologists based on science fiction

Teaching professional ethics differs considerably from teaching other subjects within a technological degree. 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 some consolidated ethical theories that they should know about, philosophical texts are often too abstract for computer scientists and engineers, and a pragmatic option is usually taken.

According to Sullins (2015), the main ethical theories relevant to digital technologies 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 (2008) 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 (2011) claims: “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.” Thus, some Ethics in Technology courses recur to science fiction 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 social robots and artificial intelligence.

Given this relevance, it is natural that instructors teaching Ethics in technological degrees are recurring to such science fiction 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. (2018) 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 more specific concerns raised by new technologies. Not only novels and short stories, but also recent movies and TV series delineate ethically-sensitive situations with considerable depth and rigor. This is the case of series like *Real Humans* and *Black Mirror*, which could trigger very elaborate and even scholarly debate, as well as films like *Blade Runner 2049*, *Surrogates* and *Robot and Frank*. Actually, the latter is being used in the platform Teach with Movies (2012) as a guide for a high-school course on Robot Ethics.

In an academic context, Iverach-Brereton (2011) reviews the roles played by robots in movies from a historical perspective, paying special attention to their degree of autonomy, and uses such fictional scenarios as a tool to make predictions about how humans may or may not accept robot integration into society. Similarly, El Mesbahi (2015) explores ethical issues related to human-robot interaction through the lens of thirty popular sci-fi movies, and presents the results of a survey about how people perceive robots in those movies and who they feel is responsible for their actions, namely the robot itself, the designer/manufacturer, the programmer or the user.

Turning to books specifically designed to teach courses in technological degrees based on science fiction, I would highlight Murphy (2018), Nourbakhsh (2013) and Torras (2018); see Figure 2. The one edited by Murphy is intended to explain key principles of artificial intelligence; the enclosed stories —by I. Asimov, V. Vinge, B. Aldiss, and Ph. K. 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."

ROBOTICS THROUGH SCIENCE FICTION
Artificial Intelligence Explained Through Six Classic Robot Short Stories
edited by
ROBIN R. MURPHY
"Concise and illuminating with few detours into engineering-speak, this book will interest a broad technology-minded audience, from future scientists to data analysts to entrepreneurs." —Publishers Weekly
ROBOT FUTURES
ILLAH REZA NOURBAKHSH
THE VESTIGIAL HEART
A NOVEL of the ROBOT AGE
CARME TORRAS

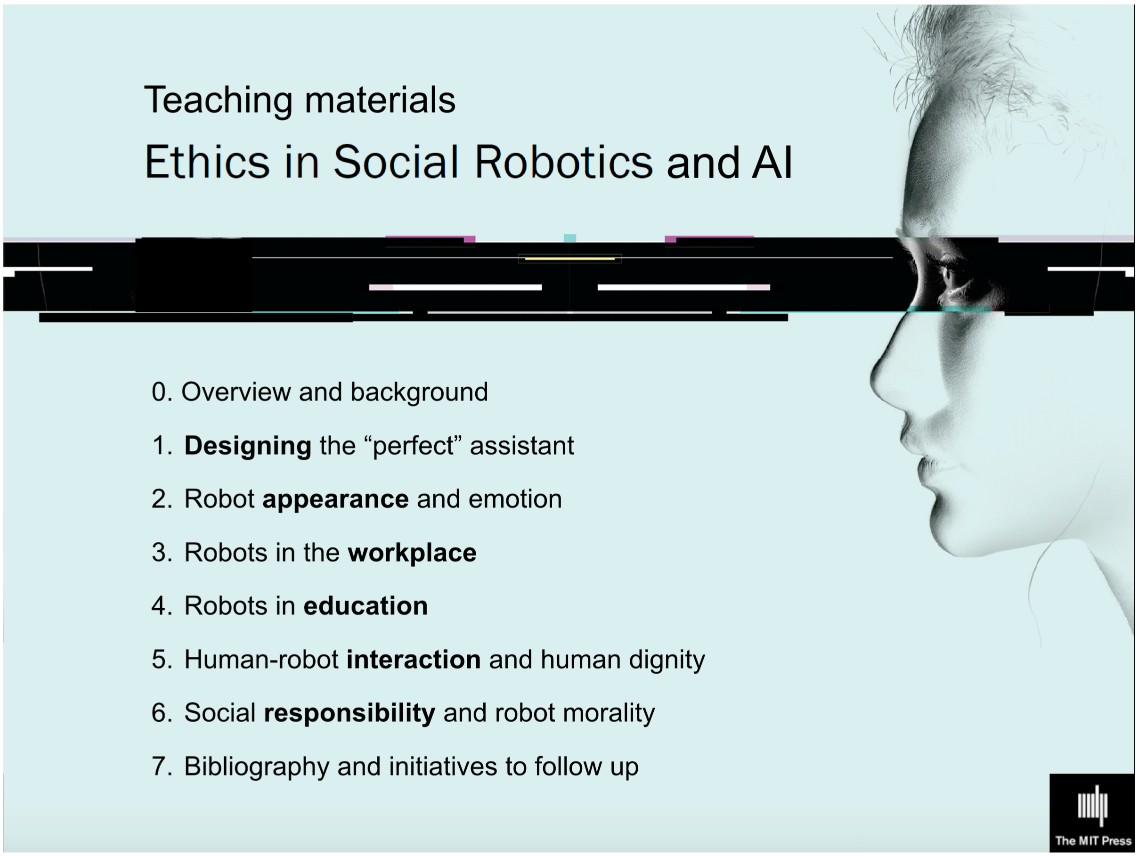
Figure 2. Three books published by MIT Press specifically designed to teach courses in technological degrees based on science fiction.

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 (2013) 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 (Torras 2016, 2019), 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* (Torras, 2018), 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 (1977): «it is the relationships that we have constructed which in turn shape us». 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.

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* (see Figure 3). 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 programming «morals» in robots.



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

Figure 3. Cover of the 100-slide presentation and chapters in the teacher's guide of a university course on Ethics in Social Robotics and Artificial Intelligence, downloadable as free ancillary materials from MIT Press website: http://mitpress.mit.edu/books/vestigial-heart

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. The book together with the ancillary ethics materials have been used to teach an entire Ethics in Technology course or some sessions 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.

5. Concluding remarks

The increasing interaction of people with all sorts of devices and 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 university degrees getting closer to the humanities, so that students become aware of possible sensitive issues they may face in their future careers and learn to reflect and discuss about them.

Philosophy, psychology, social sciences, and law are providing perspectives and prior knowledge to deal with these issues, while science fiction permits freely speculating upon potential scenarios and the role humans and machines may play in the *pas de deux* that irredeemably connects us. Along this line, educational materials based on science-fiction (SF) stories have proven very effective in engaging technology students taking ethics courses (Burton *et al*. 2018).

This growing need for practical and engaging ethics courses within Engineering and Computer Science curricula, recognized in USA several years ago, is now rapidly spreading to other countries, particularly in Europe, and to related disciplines as well. Several syllabus relying on both ethics texts and classical SF stories are being proposed and the corresponding courses will be taught at more and more universities worldwide in the coming years.

We have here 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.

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

ACM/IEEE CS Curricula (2013) Curriculum Guidelines for Undergraduate Degree Programs in Computer Science. https://www.acm.org/binaries/content/assets/education/cs2013_web_final.pdf

Borrego, M. and L. K. Newswander (2008) Characteristics of successful cross-disciplinary engineering education collaborations. *Journal of Engineering Education*, 97(2), 123-134.

British Standards Institution (2016) Robots and robotic devices. Guide to the ethical design and application of robots and robotic systems. Online: https://shop.bsigroup.com/ProductDetail/?pid=000000000030320089

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

El Mesbahi M. (2015) Human-Robot Interaction Ethics in Sci-Fi Movies: Ethics Are Not ‘There’, We Are the Ethics! Intl. Conference of Design, User Experience, and Usability, *Lecture Notes in Computer Science*, 9186, 590-598.

euRobotics project (2012) Deliverable D3.2.1 - Ethical Legal and Societal issues in Robotics. http://www.eurobotics-project.eu/cms/upload/PDF/euRobotics_Deliverable_D.3.2.1_ELS_IssuesInRobotics.pdf

European Parliament (2017) Civil Law Rules on Robotics. Online: http://www.europarl.europa.eu/RegData/etudes/PERI/2017/580862/IPOL_PERI(2017)580862_EN.pdf

ICRA Forum (2013) Robotics Meets the Humanities. http://www.icra2013.org/indexaf5b.html

IEEE Standards Association (2019) The Global Initiative on Ethics of Autonomous and Intelligent Systems. https://standards.ieee.org/develop/indconn/ec/autonomous_systems.html

IROS Forum (2018) Robotics meets the Humanities: Social Relationship, Ethics, Art and Science Fiction. http://www.ynl.t.u-tokyo.ac.jp/forum/iros2018/

Iverach-Brereton C. (2011) Learning from the Future: What Science Fiction Can Teach Us about Social Robots. *Advanced Introduction to Human Robot Interaction (AHRI 2011)*, University of Manitoba, Winnipeg, Canada.

Murphy R. (ed.) (2018) *Robotics Through Science Fiction: Artificial Intelligence Explained Through Six Classic Robot Short Stories*. MIT Press.

Nourbakhsh I.R. (2013) *Robot Futures*. MIT Press. Cambridge, Massachusetts.

RoboLaw project (2014) Deliverable D6.2 - Guidelines on Regulating Robotics. Online: http://www.robolaw.eu/RoboLaw_files/documents/robolaw_d6.2_guidelinesregulatingrobotics_20140922.pdf

Rushkoff D., Hammond R., Thomas S., Markus H. (2012) *The Tomorrow Project*. Bestselling Authors Describe Daily Life in the Future. Intel. Santa Clara.

Solomon R.C. (1977) *The Passions*. New York: Anchor Press / Doubleday.

Stephenson N. (2011) Innovation starvation. *World Policy Journal*, 28(3), 11-16. Online: http://www.worldpolicy.org/journal/fall2011/innovation-starvation

Sullins J.P. (2015) Applied professional ethics for the reluctant roboticist. In *The Emerging Policy and Ethics of Human-Robot Interaction*, edited by L.D. Riek, W. Hartzog, D. Howard, A. Moon and R. Calo, Workshop at the 10th ACM/IEEE International Conference on Human-Robot Interaction, Portland.

Teach with Movies (2012) Robot Ethics Using Clips from *Robot and Frank.* http://teachwithmovies.org/robot-and-frank/

Torras C. (2016) Service robots for citizens of the future. *European Review*, 24(1), 17-30.

Torras C. (2018) *The Vestigial Heart*. *A Novel of the Robot Age.* MIT Press. Cambridge, Massachusetts. (A teacher's guide and a 100-slide presentation are provided as free ancillary materials for instructors: http://mitpress.mit.edu/books/vestigial-heart)

Torras C. (2019) Assistive robotics: Research challenges and ethics education initiatives. *DILEMATA: International Journal of Applied Ethics*, 30: 63-77.

Veruggio G. (2005) The birth of roboethics. Roboethics Workshop at the IEEE Intl. Conf. on Robotics and Automation, Barcelona.

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

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