Teacher's Guide to **`Ethics in Social Robotics´**

based on the science-fiction novel **`The Vestigial Heart´**

by **Carme Torras**

0. Overview and Background

This guide comes as a complement to the book *The Vestigial Heart – A Novel of the Robot Age* (MIT Press, 2018) and extends its appendix, where some discussion topics and questions for reading groups are outlined. Here, academic background for the related ethical issues is provided, together with suggestions for several sessions of debate and relevant up-to-date references for further reading.

The materials are intended for teaching "ethics in social robotics" at the university level, especially in technological degrees such as computer science and engineering, but also in philosophy, psychology, political science, cognitive science, and linguistics, which all have ethics-related topics in their curricula.

The structure of this guide is as follows. After covering the motivation and preliminaries on roboethics in this introductory section, the remaining six sections run in parallel to the appendix in the book, each being devoted to an ethics theme: robot design, appearance and emotion, robots in the workplace, in education, human-robot interaction, and social responsibility. Since the book appendix includes suggested readings from the novel and four questions on each theme, the corresponding sections here start with quotation highlights from such readings that exemplify the issue addressed, followed by an overview of the scholarly previous work on each ethical question intertwined with hints to trigger further debate. Finally, in the last section, the available texts on roboethics are briefly discussed, as well as some ongoing initiatives for interested people to follow up.

0.1. Motivation: The Need for an Ethics Debate on Social Robotics

In the coming years we will find ourselves increasingly interacting with robots as part of our daily lives. Robots will attend to elderly and disabled people, perform household

tasks, act as support teachers, assistants in shopping malls, receptionists, guides at trade-fairs and museums, and even as nannies and playmates. Robots will perform service tasks in smart cities as well, such as logistics, cleaning and recycling, surveillance and environmental monitoring; and in factories they will not just be caged in production lines but they will also work in close collaboration with humans.

The International Federation of Robotics periodically delivers statistical data that substantiate the impressive growth of service robots and personal assistants, which will presumably keep accelerating in the near future.

The speed at which robotics technology develops outstrips the ability to establish guiding principles for their use, leaving robot manufacturers and programmers as unintended policy makers influencing society, and pushing consumers to seek suitable information that would enable them to make proper choices. This state of affairs calls for the development of practical ethics guidelines for robotic engineers, as well as materials to teach roboethics to university students of related disciplines, which could eventually increase the awareness of society at large.

Service and assistive robots pose a much wider range of ethical issues than their industrial predecessors and other machines, as they enter domains previously exclusive to humans, such as decision-making, feelings, and relationships. In order to regulate their uses for public benefit, it becomes of upmost importance to predict how increasingly close and frequent relations with robots will influence individual identity, society and the future of humankind.

However, significant methodological difficulties arise when formally undertaking such prediction [Ballesté and Torras 2013]. Unforeseen uses always crop up for any kind of technology developed, as in the case of cell phones, initially intended for commercial interactions. Technological development cannot be studied outside its sociocultural context, and the limitations of language to describe the future should also be kept in mind; quoting Heidegger, it is «through technique that we perceive the sea as navigable».

Given the difficulty of predicting how a technological society will evolve, a reasonable way out is to imagine different possible future scenarios and encourage debate on their advantages and risks. Several initiatives have involved science-fiction writers in trying to outline varied and consistent future scenarios; an example is the Center for Science and the Imagination.

Because of my research on assistive and collaborative robotics, I often attend prospective, brainstorming meetings where we try to foresee what we may be ready to offer in five or ten years' time, from a technological perspective. I became progressively concerned about the social and ethical implications of the potential innovations we were discussing. This concern, given the methodological difficulties mentioned above, encouraged me to try my hand at fiction, and in the novel *The Vestigial Heart,* I imagined how being raised by artificial nannies, learning from robot teachers and sharing work and leisure with humanoids would affect the intellectual and social habits of future generations, their feelings and relationships, enhancing or spoiling them depending on viewpoint. The novel's *leit motiv* is a quotation from the philosopher R.C. Solomon: «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 robot companions, if they are to pervade our lives.

Following a suggestion by MIT Press Editor Marie L. Lee, I undertook the structuring of ethics themes around the situations appearing in the novel. It turned out that, since I originally resorted to fiction precisely to unravel the social and ethical issues that might appear when robot assistants were in common use, most of the themes related to human-centered robotics are well-illustrated in the novel.

Incidentally, the non-covered ethics topics are those most often addressed in the currently available courses, projects and texts on robot ethics – namely military aspects, legal regulations, surgical robots, as well as highly speculative issues like the possibility of endowing robots with consciousness and morality. There seems to be a gap in the available roboethics materials regarding *the practical concerns of engineers*

and laymen about these assistive robots now being designed to share our everyday lives in the near future. This teacher's guide aims at filling this gap.

0.2. Preliminaries on Roboethics

The term Roboethics was introduced by G. Veruggio at the beginning of the century and refers to the 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. Subtle distinctions are often made between human ethics applied to robotics, codes of ethics embedded in the robots themselves (sometimes named "machine ethics"), and ethics that would emerge from a potential future consciousness of robots [Veruggio *et al*. 2011]. We will here concentrate on the first and touch partially on the second, leaving the third one to professional philosophers.

Computer ethics, which studies and analyzes the social impacts of information and communication technologies, is a related subfield of applied ethics with a much longer trajectory. It addresses issues like privacy, intellectual property, safety, reliability, autonomous and pervasive technologies, vulnerable groups, and professional ethics, which fall also within the scope of Roboethics. However, the fact that a robot is not only equipped with a computer, but also with sensors and actuators allowing it to act in the real world, raises another whole set of questions mainly derived from its close-to-human functionalities and its degree of decision-making autonomy. As Nourbakhsh (2013) suggests, robots can be thought of as the interface of the digital world with the physical world, and therefore they carry issues previously restrained to digital information to the reality of physical objects and environments.

There are many ethical theories relevant to robotics. Sullins (2015) briefly surveys 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).

Following our pragmatic option and since no single theory is appropriate for addressing all ethical issues arising in the design and use of robots, we adopt a hybrid approach here. Such hybrid ethics are 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). In developing the teaching material below, we will acknowledge and apply the ethics theories that seem more suited to each particular case.

References

Ballesté F. and Torras C. (2013) Effects of human-machine integration on the construction of identity. In *Handbook of Research on Technoself: Identity in a Technological Society*, edited by R. Luppicini, IGI Global, pp. 574-591.

Nourbakhsh I.R. (2013) *Robot futures*. MIT Press.

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.

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

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

1. Designing the "perfect" assistant

1.1. Highlights from *The Vestigial Heart*

Chapter 1, page 5:

Alpha+ has arranged everything to perfection. He [Dr. Craft] would trust Alpha+ over and above his mother, if she were still alive. Or, it goes without saying, his daughter or his wife. The robot was already a good servant, but since it had the neuroaccelerator installed it is learning at a vertiginous speed, and in a few days has adapted to him like a tight glove.

[..] A good choice of stimuli, that's the secret to wellbeing. Let's forget about self-help implants and other neuropsychological devices, we can't change man or turn his brain upside down, we can't modify even the smallest reaction. Let's accept that. The only way forward 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. Too simple, they said. How short-sighted! One must understand man, each man, in order to be able to activate the right resources at the right time. This was the difficult part: 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 or how much sugar to put in his coffee, the whole idea would go down the drain.

Chapter 1, page 7:

So many prostheses for everything nowadays, and there's not even one for his handicap. Damn the LED contact panels! What he wants is a creativity prosthesis. Or an assistant, it doesn't matter; 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. Now the net has placed him on the massager at the side of the bathtub and a series of cushioned rolls and

strategically placed heat sources are drying and massaging him from head to toe. A brain massage, that's what he needs.

Chapter 5, page 25:

[ROBco:] ... but my PROP [Leo] does not heed alarm signals. In critical situations, he interferes directly and changes priorities. This goes against all the factory specifications and security regulations. He must know what he is doing, because this is his field, and they say he is one of the best, but he often skips maintenance, and for several months he has ignored pending updates and expansion notifications.

Chapter 7, page 40:

If he could extract the creative potential of Mar'10 and combine it with the wise, mature loyalty of Gatew ... that would be a cutting-edge ROB.

1.2. Ethical Background and Discussion

The traits attributed to a "perfect" assistant vary largely among cultures, as well as among individuals. Thus, some robotic assistants are designed targeting specific populations, whereas other more generic designs are left open to tuning by end-users. This tuning can be purposeful or result from automatic adaptation. From an entrepreneurial viewpoint, ***Doctor Craft*** argues for highly-adaptable robots that would fit their owners like a glove, covering all of their needs and hopefully maintaining them in a permanent state of wellbeing, but as a user he requests a hypothetical assistant to stimulate his thinking and to inspire him to behave differently than usual. Similarly, ***Leo*** presumably adheres to more strict criteria (e.g., as regards to safety and maintenance) in his professional design activity than when he tunes his robot as a user.

Robot designers and programmers are becoming aware of their influencing capacity on the customers' way of living, and that early design stages make a greater impact than latter development ones. If traditional design ethics was mainly concerned with functionality, sorting out responsibilities for failures and assessing measures for risk prevention, nowadays it takes into account the impacts on users' moral decisions and actions, and on the quality of their lives. As the mediating role of digital technologies

has been recognized [Verbeek 2008], design with embedded ethics has begun to be discussed. Ingram *et al*. (2010) have registered a very generic code of ethics for robotics engineers, which among other principles includes the responsibility to keep in mind at all times the wellbeing and expectations of customers and end-users. But, as mentioned, these expectations may vary a lot.

To discuss potentially desirable features of an artificial assistant, Peltu and Wilks (2008) searched inspiration in the virtues attributed to the Victorian lady's companion. Among those applicable to personal robots, we highlight the following: able to distinguish its owner from other people, animals and things; able to recognize its owner's emotions and intentions; behaving in a predictable and dependable way; protective and supportive of the user in the handling of information and communication with other people; polite but firm in the owner's interest; having a model of its own capabilities; operationally reliable and requiring neither much effort from the owner to use nor special maintenance.

Wish lists like this one have not only ignited the imagination of researchers, but also have fostered debate and ultimately shaped some robot design guidelines. For example, Knight (2014) provides some smart social design considerations that, in her words, may help avoid unnecessary policy friction in the future. Among them, I would highlight the implementation of safeguards to ensure robots augment human experiences rather than increase social barriers, and designing robots to be intentionally machine-like, since humans view robots as agents and react to them socially.

In 2012, a report from South Korea Robot Ehtics Charter was released, consisting of three parts devoted to regulations for manufacturers, owners/users, and robots. The first part lists seven very pertinent manufacturing standards, which can be summarized as: i) limit robot autonomy by making it always possible for a human to assume control; ii) guarantee user and community safety; iii) minimize the users' risk of any psychological harm such as antisocial or sociopathic behaviors, depression or anxiety, stress, and addictions; iv) clearly identify the product and protect it from alteration; v)

protect personal data; vi) ensure that robot actions are traceable at all times; and vii) make designs ecologically sensitive and sustainable.

Additionally, Van der Loos (2007) brings about the important principle of transparency, whereby software developers must pair each new layer of complexity in robot behavior with a corresponding communication layer for conveying the intention of those behaviors to the surrounding people (and robots) through, for example, gestures, voice and context. This emphasis on communication is supported by a field study carried out by Dautenhahn *et al*. (2005) on people's preferences as regards to assistant robots, where humanlike communication was largely prioritized over humanlike behavior and appearance.

The code of ethics focused on human-robot interaction proposed by Riek and Howard (2014) extends and refines some of the characteristics already discussed. Transparency in robot behavior is extended to transparency in programming as well as to predictability of future robot moves; real-time status indicators are suggested to increase trustworthiness and, depending on the type of users, kill switches could enhance their perception of safety. Special attention is devoted to deception from improper Wizard-of-Oz use, which will be discussed later under Question 3.C.

Question 1.A – Should public trust and confidence in robots be enforced? If so, how?

Given the understandable concerns of some people about the rapid pace of technological change and the role robots could play in our future society, surveys are periodically conducted to gauge public opinion about robots, and to assess the extent to which people will accept robots performing certain functions. The Special Eurobarometer 427 on Autonomous Systems (2015) is one such survey requested by the European Commission. A noticeable conclusion is that while personal experience with robots is rising, the proportion of respondents expressing a positive view has declined since 2012, from 70% down to 64%. The evolution of people's attitude towards robots is an important issue debated at roboethics forums.

Principles of robotics, issued by the Engineering and Physical Sciences Research Council of the UK, start by saying that the realities of robotics are still relatively little known to the public and that steps should be taken «to ensure that robots are introduced in a way that is likely to engage public trust and confidence», so that this technology is integrated into our society to the maximum benefit of all its citizens and proactively heads off any potential unintended consequences. Then, it is advised that «we, roboticists, take responsibility for our public image and demonstrate that we are committed to the best possible standards of practice». As an example, many people are frustrated when they see outrageous claims in the press that could be corrected by a simple word to the reporters, and «we should commit to take the time to contact them».

Despite this viewpoint being widely shared, it is not so clear how such a trustworthy public image could be more generally conveyed. Nourbakhsh (2010) argues that roboticists tend to employ an inadequate rhetoric to justify the interest of some robot applications for society. They often recur to value hierarchy (i.e., robots don't need to be perfect, but just do better than the current way of accomplishing a task) and semantic inflation (i.e., describe robot cognition with loaded terms that contrast with the often prosaic aspect of the robot), without providing the public with the knowledge they need to elucidate their legitimate concerns (e.g., safety, undesired side effects). Thus, Nourbakhsh claims that roboticists should employ a language for communication that empowers the audience to make the most appropriate possible decisions (e.g., characterizing a robotic assistant for the elderly in terms of backdriveability of its mechanism in case of computational malfunction). Since perceiving an innovation as beneficial or not often depends on expectations regarding its future impact, and non-experts have trouble disambiguating short-term from long-term consequences, he advocates for adding a section to robotics publications that would explicitly describe the short-term (five years and less) and long-term (ten years or more) implications of the new result.

Question 1.B – Is it admissible that robots be designed to generate reliance?

Veruggio (2005) raised the issue that «addiction to robots could be more dangerous and disrupting than to TV, internet, and videogames». A reason for this is that robots can cover a much wider range of areas in our daily life than other technologies. But addiction goes farther than the technological dependence inevitable in technified societies. For instance, in benchmarking human-robot interaction, the question of where the boundary lies between comforting exercises and addiction to robots in elderly groups often arises [Espingardeiro 2015].

As mentioned earlier, the South Korea Robot Ethics Charter lists among its design and manufacturing standards to minimize the user's risk of addictions. But this may sound contradictory to what Oliver (2015), Scientific Director of R&D for Telefónica, claims: «We have to understand that technology is designed to be addictive, otherwise companies would not make money. There is no point in being naive or innocent about this: a lot of research and preliminary work goes into it. [..] 78% of adults in the United States regard themselves as nomophobic, i.e. they get anxious and experience physical symptoms if they do not have their mobile handy. This should give us food for thought».

How robot designers can cope with the sometimes opposite interests of companies and users is a classical question open to debate. Some would argue that business competition and public education would result in products satisfying them both [Roberts 2001].

Question 1.C – Should the possibility of deception be actively excluded in the design of robots?

The risk of deception in the social deployment of robots is high and takes many forms depending on context. To name but a few, elderly people may be deceived into believing that their robot assistants care about them, children may have the induced illusion that robot toys have mental states and emotions, and the general public may be deluded to think that robots are truly intelligent and have intentions. Of course,

some cases are more morally reprehensible than others [Matthias 2015], but the general agreement is that robots should be designed in ways that do not impersonate human agency by attempting to mimic intentional states.

The paradox, especially in the case of humanoid robots, is that their design conveys human attributes, thus fostering this deceit problem. Moreover, as Breazeal (2015) adverts, «give hearing and voice to a robot and people expect it to be intelligent». Even if users know they are talking to a machine, they tend to respond as if it has some sort of consciousness and sense of purpose.

All in all, the Principles of robotics mentioned earlier state that «robots should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be transparent».

Question 1.D – Could robots be used to control people?

Some researchers have expressed a fear that society has become too complacent about the potential of digital technologies to be used to heighten surveillance and control over citizens. The opinion that «if you have nothing to hide there is no need to be concerned» is spreading quickly. However, a great deal of such information has been used to repress people and political movements, so it seems over-confident to imagine that no regime would ever misuse data within your, or your data's, lifetime [Peltu and Wilks 2008].

But this is not just a matter of privacy regarding the data a user voluntarily uploads. Not only robot assistants may share personal material without the user knowing, but information may flow the other way too, influencing personal choices and ultimately manipulating people. This is what Lowe (2010) refers to as «the watching eye and punitive hand of the state». Robots may enforce certain habits and values on the user, the key questions being who decides which these should be and whom they would benefit: the user, society at large, or a particular group of people. If it is the user that, for example, wants to follow a diet, he himself may tune the robot to distract him from eating between meals, or to act as a kind of Jiminy Cricket by reminding him how

ashamed he will be later on. Similar behaviors may be programmed into robots to encourage certain healthy habits in their users with an eye to reduce the social medical expenditure, but this programming can likewise be used to increase the economic benefits of some companies or to favor the political interests of a party or state.

Borenstein and Arkin (2016) refer to this robot tactic of subtly influencing its user as "nudging". Following Thaler and Sunstein (2008), they distinguish three degrees of such tactic: weak paternalism (preserving an individual's wellbeing as presumably he would like to), libertarian paternalism (molding human behavior toward more productive ends, without blocking or fencing off choices), and strong paternalism (protecting someone against their voluntary choice by legally implementing security measures). Three design pathways are then envisaged: "opt in" (the user selects preferences), "opt out" (there is a default setting that the user can modify) and "no way out" (certain alarms cannot be disabled or some limits cannot be surpassed).

These authors pose the interesting question of whether «it is ethically appropriate to deliberately design nudging behavior in such a way so that it increases the likelihood that the human user becomes ''more ethical'' (however that is defined)». The first example they mention is set in a private context (e.g., redirect the user attention from completing work to a child that has been sitting along watching television for a long time), but far-reaching implications in the public domain (e.g., promoting social justice) are next envisaged: «a robot could access its owner's schedule and then nudge her to be involved in adult literacy campaigns when "free time" is available or respond to an emailed emergency charitable donation request when that request is deemed legitimate». Now, if designing robots that enforced social justice were both technically feasible and ethically acceptable, wouldn't there be a moral imperative to build them? We will come back to this issue restricted to the domains of education and healthcare in Sections 4 and 5, respectively.

One could reasonably argue that the above concern is not specific of robots, since other devices such as cellular phones and intelligent watches have some nudging capabilities, as well. However, it is worth emphasizing that the potential of personal

robots is immensely higher, because their autonomous motion permits following and monitoring the user, and their compelling physical presence is much more persuasive [Li 2013].

Even if performed in the interest of the user, nudging can be perceived as overly intrusive and annoying, thus running a high risk of angering people, especially bad-tempered personalities. ***Doctor Craft*** is such kind of user, and this situation is already illustrated in the first scene of the novel, when he roars to his robot ***Alpha+***: «Get off me, you confounded beast» and gives it a shove as it is trying to wake him up. Thus, nudging effects depend a lot on the user and the circumstances, which has to be carefully taken into account at design time.

In sum, while some ethical guidelines for the professional practice of robot designers and programmers have been established, a consensus on what constitutes an ethical robot behavior is far more difficult to reach, if not impossible in generic terms. Hopefully, when facing a particular case, roboticists will find the discussion on open design issues in this section useful to trigger their creative thinking and eventually come up with criteria valid for that specific case.

References

Borenstein J., Arkin R. (2016) Robotic Nudges: The Ethics of Engineering a More Socially Just Human Being. *Science and engineering ethics*, 22(1): 31-46.

Breazeal, C. (2015) Jibo Is as Good as Social Robots Get. But Is That Good Enough? *IEEE Spectrum*. http://spectrum.ieee.org/robotics/home-robots/jibo-is-as-good-as-social-robots-get-but-is-that-good-enough

Dautenhahn K., Woods S., Kaouri C., Walters M.L., Koay K.L., Werry I. (2005) What is a robot companion-friend, assistant or butler? *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, 1192-1197.

Espingardeiro A. (2015) Social Assistive Robots, Reframing the Human Robotics Interaction Benchmark of Social Success. *Intl. J. of Social, Behavioral, Educational, Economic, Business and Industrial Engineering*, 9(1): 377 - 382.

Ingram B., Jones D., Lewis A., Richards M., Rich C., Schachterle L. (2010) A code of ethics for robotics engineers. *5th ACM/IEEE Intl. Conf. on Human-Robot Interaction*, 103–104. Registered in the Codes of Ethics Collection, Illinois Institute of Technology, http://ethics.iit.edu/ecodes/node/4391

Knight H. (2014) How humans respond to robots: building public policy through good design. *Brookings Report*. http://www.brookings.edu/research/reports2/2014/07/how-humans-respond-to-robots

Li J. (2013) The nature of the bots: how people respond to robots, virtual agents and humans as multimodal stimuli. *15th ACM Intl. Conf. on Multimodal Interaction (ICMI '13)*, New York, 337–340.

Lowe W. (2010) Identifying your accompanist. In Wilks Y. (Ed.) *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues*, Natural Language Processing, 8: 95–100.

Matthias A. (2015) Robot Lies in Health Care: When Is Deception Morally Permissible? *Kennedy Institute of Ethics Journal*, 25(2): 169-162.

Nourbakhsh I. (2010) The Rhetorics of Robotics. Unpublished manuscript. https://sites.google.com/site/ethicsandrobotics/home/living-archive/reading-and-reading-questions/nourbakhsh/RhetoricofRoboticsIRN.pdf

Oliver N. (2015) Nothing in excess; including technology. *Barcelona Metropolis*, 86, 42-44. http://w2.bcn.cat/bcnmetropolis/wp-content/uploads/2015/06/BMM96.pdf

Peltu M., Wilks Y. (2008) Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues. *Oxford Internet Institute, e-Horizons Forum Discussion Paper*, 14.

Riek L.D., Howard D. (2014) A code of ethics for the human-robot interaction profession. *We Robot Conference,* University of Miami.

Roberts R. (2001) *The Invisible Heart – An Economic Romance*, MIT Press.

Special Eurobarometer 427 on Autonomous Systems (2015) http://ec.europa.eu/COMMFrontOffice/PublicOpinion/index.cfm/ResultDoc/download/DocumentKy/65859

Thaler R.H., Sunstein C.R. (2008) *Nudge: Improving decisions about health, wealth, and happiness*. New Haven: Yale University Press.

Van der Loos H.M. (2007) Ethics by design: A conceptual approach to personal and service robot systems. *Roboethics Workshop at the IEEE Intl. Conf. on Robotics and Automation,* Roma.

Verbeek P. (2008) Morality in Design: Design Ethics and the Morality of Technological Artifacts. In Vermaas P.E., Kroes P., Light A., Moore S.A. (Eds.): *Philosophy and Design*, 91-103, Springer.

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

2. Robot Appearance and Emotion

2.1. Highlights from *The Vestigial Heart*

Chapter 9, page 53:

At the end of the party Lu [adoptive mother] gave me [Celia] my very own robot just for me. Yes, a robot, I know it will be difficult for you to imagine. So you have an idea, it's like the ones from Star Wars, but, along with legs, it has four wheels for when it wants to move quickly, and it doesn't have a face. Well, 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. I'll have to learn to use it, even though it does a lot of things on its own already. It will go with me everywhere; to start with, we'll go to school together tomorrow. I didn't know, but everyone has their own robot here; it's like us having a wallet or a calendar, but much more sophisticated, because it has a large memory and can solve problems for you. Lu's is called ROBul, and it's been hidden all this time so as not to scare me. I don't understand why. I've been more shocked by the kids, and even some things Lu does, than by ROBbie. For the robot, everything follows a series of rules, it'll never surprise me with anything inappropriate.

Chapter 10, page 58:

He [Leo] is not sure what drew him to this realistic mechanical baby, if it's the grotesque expression, the diapers it doesn't need or the fact that it bears the logo of Bet's company. First they brought out those practical little dogs that didn't need to poop or pee, and then they started mimicking wilder and wilder animals, until they got to man. 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? And, best of all, that can be turned off and shut in the cupboard when it gets whiney and tearful? Well no sir, the product didn't take off, almost certainly because it's too much like the real thing, déjà vu.

Chapter 12, page 79:

Celia stops for a moment, touched by the words, and 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. As they start moving again, 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. And what does it matter that he doesn't have eyes, people don't look at each other anymore anyway.

2.2. Ethical Background and Discussion

People's attitude towards robots and the factors influencing it have been the subject of numerous studies (e.g., see [Li *et al.* 2010] and the Special Eurobarometer 427 on Autonomous Systems (2015) in the preceding section). Appearance has been shown to play a prominent role: the more anthropomorphic the robot, the more positive and empathetic the human response [Riek *et al.* 2009]. This positive relation has been demonstrated even at the neurological level through fMRI recordings [Krach *et al.* 2008]. However, the relation doesn't grow unlimited; on the contrary, a point is reached where excessive similarity of the robot to a human causes distress and provokes a sudden repulsion; this is known as the "uncanny valley" effect and will be discussed under Question 2.C below. ***Celia*** feels attached to her robot ***ROBbie*** because of its loyal, trustworthy and predictable behavior, which is enforced by its undeceiving machine appearance. ***Leo*** realizes that a too-close similarity to a human being can doom a robot product.

Two reasons are usually put forward to design robots with humanoid shape. One is functional, since such robots can operate interchangeably with humans in their very same environments and making use of the same tools, machines, vehicles, etc. Thus, there is no need to predefine or remodel workspaces for them. The other reason is precisely to be better accepted by humans, although as mentioned, there are limits to this.

Note that anthropomorphism refers not only to shape but to emotional impersonation as well, including attention, intentionality and, above all, expressiveness. Many experiments have been conducted to explore the minimal requirements for effective human-robot social interaction, in terms of facial expressions [Bruce *et al.* 2002], full body postures [Beck *et al.* 2012], as well as other progressively less anthropomorphic modalities, such as sounds and eye color [Häring *et al.* 2011]. By combining human-like, animal-like, and robot-specific social cues on a mildly humanized robot, Embgen *et al.* (2012) showed that robotized abstractions of emotion display could in fact be an alternative to complex facial expressions (difficult to implement in robots), leading to human-robot interactions not mimicking human-human ones.

This permits avoiding the risk of deception in contexts where the functionality of the robot does not require it to have anthropomorphic features, without diminishing its social communication abilities.

Question 2.A – How does robot appearance influence public acceptance?

Coeckelbergh (2009) argues that the impact robots have on us depends on how they appear to us, not on what they "really" are and their true cognitive abilities. Thus, he advocates taking seriously the ethical significance of appearance and turning to a philosophy of interaction in order to establish an ethics of appearance and human good. As we will further explain in Section 5, such ethics entails listening to people's experience and using our moral imagination to find out possibilities of living with robots that enhance human flourishing and happiness. In a few words, this is an open-minded, bottom-up approach that, instead of setting up moral limits to the design of robots, focuses on human-robot interactions and the way these may enrich our emotional life in a possibly different and complementary way to human-human relationships.

Along this line, Duffy (2003) asks: «Similar to the argument to not constrain virtual reality worlds to our physical world, are we trying to constrain a robot to become too animalistic (including humanistic) that we miss how a robot can constructively contribute to our way of life?» Leaving this question open, he concludes that

anthropomorphism provides us with very powerful physical and social features that will no doubt be implemented to a greater extent in assistive robots as they ease communication with users.

In order to explore alternative robotic morphologies that could enrich people's daily interactions, Sirkin and Ju (2014) have robotized some everyday objects to appropriately respond to human intentions and emotions, and Sabanovic *et al*. (2014) have proposed innovative prototyping methods for designing socially situated embodiments.

A more generic ethical consideration related to robot appearance, which almost goes without saying, is the need to avoid sexist, ableist, racist and ethnic robot morphologies and expressiveness in the design and programming of robots.

Question 2.B – What are the advantages and dangers of robots simulating emotions?

There is no doubt that emotion expression by a robot plays an important role in social, face-to-face interactions with people [Breazeal 2003]. What is not so clear are the advantages and dangers that such expressiveness entails, which of course depend on the particular circumstances.

In a search and rescue setting, Moshkina (2012) showed that nonverbal expressions of negative mood and fear by the robot improved the participants' compliance with its request to evacuate, causing them to respond earlier and faster. Even in the absence of explicit requests, just the "nervousness" of the robot had the positive effect of making people more alert to any unfavorable changes in the surroundings.

On the other hand, in companionship settings, Cowie (2014) identifies deception as an important danger of affective simulation. He argues that deception infringes autonomy, because misinforming a person about the alternatives that are open prevents him or her from choosing rationally between them. An obvious illustration is when the robot companion uses facial and vocal gestures that give an impression of caring. That may help the robot in its intended function, but it is a problem if the user

drifts into assuming that it will show other kinds of caring behavior, and relies on it for help that it cannot actually provide.

Several situations need to be distinguished here. The most sensitive case is that of robots designed to take care of vulnerable users, like children and elderly people without full adult judgment, and special attention should be paid to the design of such robots. At the opposite extreme, some adults gladly welcome predictable relational artifacts as substitutes for the often resistant and difficult-to-live-with human beings.

There is a difference between simulating affection and showing emotional intelligence. The latter entails capturing the emotional state of the user and acting accordingly, which can be very handy in some healthcare situations. This will be addressed in Section 5, but let us mention that the danger here is that the perceptual competence of robots can never be fully certified, and as a result, people will be subjected to actions that they do not deserve, or will not receive responses that they ought to. The problem is not new. The classical example involves 'lie detectors'. Despite widespread belief in their powers, they were actually much more likely to stigmatize the innocent than to pinpoint the guilty [National Research Council, 2003].

Note that the potential of robots to display affection goes well beyond the text, speech and visual messaging offered by devices such as cellular phones or tablets; not only do robots have mobility to follow users wherever they go, but they also can even physically provide emotional support during a stressful situation (e.g., hugging or providing physical contact such as a pat of support). Turkle (2007) alerts that, although the robot is only expressing a simulated emotion, the feelings it evokes in people are real and may be strong. Let us close this section with an intriguing quotation from her: «In the culture of simulation, authenticity is for us what sex was to the Victorians: taboo and fascination, threat and preoccupation.»

Question 2.C – Have you heard of/experienced the "uncanny valley" effect?

Precisely the sudden perception of a lack of authenticity may cause the repulsion towards highly human-like machines that has been termed the "uncanny valley" effect.

This name follows from the shape of the curve representing human attitude as a function of robot anthropomorphism, which grows steadily up to a point where it falls down into a profound valley. Just picture the reaction triggered by the fictional robot WALL-E and compare it to that produced by humanoids such as Saya, developed at the University of Tokyo, or the initial geminoids, developed at Osaka University, which look almost human but not quite, causing a creepy impression.

Geminoids' designer H. Ishiguro is a pioneer in building robots that look like actual people, having even built a teleoperated replica of himself, among others. His claim is that, through evolution, our perceptual system has become highly sensitized to human-like appearance and behavior, eliciting an incredible range of interpersonal responses from each other; therefore anthropomorphism is a *sine qua non* to study human-robot interaction if we would like to achieve the best possible results, since subconsciously and immediately people know how to interact with robots of such form [MacDorman and Ishiguro 2006]. Following this view, the uncanny/creepy impression can help understand what human qualities robots are missing and thus trigger research to improve their communication abilities.

The "uncanny valley" effect, with its origins in Greek philosophy, has been widely discussed in psychoanalytical literature. In the robotics context, the concern is that people may be less willing to engage in interaction with such quasi-human but somehow repulsive robots. Since there is evidence that anthropomorphism can help robots to accomplish some tasks by eliciting desired behaviors from their human partners, as mentioned when discussing Question 2.B, many studies have been devoted to determining the features or lack of them that provoke such effect.

Most analyses explore progressive anthropomorphism in robot appearance, motion quality and interactivity, and recently moral values and authenticity are beginning to be considered as well. Rather than a given feature or a combination of them, what seems to trigger repulsion is the lack of coherence between the different elements that mediate human-robot communication (appearance, expression, language, speech, gestures, posture, motion, responsiveness to attentive and motivational cues,

interaction speed, turn taking in dialogs, synchronization, etc.). Any discordant element can make users feel perturbed or cheated.

Złotowski *et al.* (2015) propose to explore the uncanny valley effect the other way around, not by trying to reach the human level starting from a machine, but rather by studying humans that are perceived as lacking some human qualities. For example, Cole (2001) reports that patients who have reduced facial expressiveness caused by Moebius Syndrome or Parkinson's disease find it hard to capture the interest of others or to join in a conversation. Investigated in one way or another, it seems clear that the uncanny valley effect decreases as the user becomes more familiar with the anthropomorphic robot, and in some cases it only occurs at the very first stages of interaction.

Human-like robots raise much higher expectations regarding their capabilities compared to robots with machine-like appearance. Determining under what conditions the anthropomorphizing of machines is justified and under what conditions is unjustified is, from an ethics viewpoint, a key question to be answered in each particular case.

Question 2.D – Should emotional attachment to robots be encouraged?

Although anthropomorphism and emotionality may favor human attachment to robots, they are by no means indispensable requirements. The book edited by Wilks (2010) provides examples of how people get attached to very simple devices, leading to such irrational behaviors as refusing to board a plane because this would cause their Tamagotchi to die, or giving their bed to a doll so it can have a good night's sleep.

Aliveness seems to be the key factor for the development of affective ties with a mechanical creature, according to Bartneck *et al.* (2007). In an experimental study carried out by these authors, adult participants faced the ethical dilemma of switching off a robot they had been playing Mastermind with in a cooperative way, so that all what the robot had learned would be erased from its memory. This is a similar situation to those appearing in the movies *2001: A Space Odyssey* and *Robot & Frank*.

The results showed that participants hesitated three times as long to switch off an agreeable and intelligent robot as compared to a non-agreeable and unintelligent one. The authors hypothesized this was due to the extent the robot was perceived as a living creature, which would strongly correlate with its displayed intelligence, as a subsequent study indeed confirmed [Bartneck *et al.* 2009].

Vulnerable populations appear to be most prone to feel emotionally attached to robots. In order to objectively study attachment, a way to evaluate emotional content of users' experiences is needed. Norman (2004) categorizes experience episodes into three dimensions: visceral (i.e., first impression based on appearance), behavioral (i.e., appraisal of functionality, satisfaction of needs, and usability), and reflective (i.e., situated reasoning on the basis of past experiences and with a view to future actions). This categorization was used by Weiss *et al.* (2009) to study the attachment of children and adults to the robotic dog Aibo. They found that a visceral impression and a short-time behavioral interaction were not sufficient for adults to form an emotional attachment to the robotic dog, although it provided an indication of whether they were heading towards positive attachment or rejection in the reflective phase. On the contrary, children showed strong attraction on the visceral level, and positive emotions (like curiosity and fascination) resulted in a patient and tolerant interaction on the behavioral level. In their responses to a questionnaire, it became clear that children attributed cognitive abilities and emotions to Aibo, leading them to empathize with the robotic dog and rapidly developing an emotional attachment toward it. Likewise, other vulnerable groups such as elderly people with mild dementia tend to easily develop an emotional attachment to socially interactive robots, as will be discussed in Section 5.

Now, what are the benefits and dangers of emotionally attaching to robots? Should such attachment be enforced, discouraged, or be simply left to the users' will? Not only there is no consensual answer to these questions, but also some authors maintain quite opposite stances. For instance, Levy (2010) sustains that it is completely normal that people fall in love with artificial companions, whereas Bryson (2010) argues that machines should always be just servants that you can switch off whenever you like.

Some potential benefits of attachment (e.g., making robot nudging more compelling, and providing company/sex to adults with full judgment who deliberately make this choice) have already been mentioned, and others arising in a therapeutic context will be examined in Section 5; here we concentrate on the envisaged dangers.

The main one is the user's social isolation, which can derive from family and friends eluding their responsibilities once the care activities are covered by the robot, or be caused by the seductions of the robot itself, leading to the so-called 'lotus eater' problem [Cowie 2014]. This refers to the risk that easy attachment to a robot would erode the person's motivation for engaging with human beings, who are not always emotionally pleasant. In the case of children this can be especially harming, since reduced contact with family and peers could seriously disrupt their normal development, preventing them from learning to empathize, for example, as will be extensively debated in Section 4.

Even for fully-conscious adults, robot attachment may be good in the short-term by making them feel confortable, but bad for their long-term personal fulfillment. From a broader perspective, we could ask ourselves whether living and engaging with robots will be so easy that human relationships will be discouraged because they would just seem too hard [Turkle 2007].

A final note of caution: beyond the discussion of whether robot designs should encourage or discourage the formation of emotional bonds, roboticists should be aware that some bonding will be inevitable regardless of the morphology of the robot [Riek *et al.* 2009].

2.3. Revisiting Issues

Besides the controversies considered in this section, the readings from Chapters 9 and 12 of *The Vestigial Heart* touch on two issues discussed in the preceding section: transparency (1.C) and trust (1.A), respectively. ***Celia*** likes that ***ROBbie*** has a more predictable behavior than her classmates and her adoptive mother, since it has to

follow rules and can't shock her with nonsense. Moreover, she feels protected by the robot, which she sees as a faithful companion that she can trust.

References

Bartneck C., Van Der Hoek M., Mubin O., Al Mahmud A. (2007) Daisy, Daisy, give me your answer do!: switching off a robot. *ACM/IEEE Intl. Conf. on Human-Robot Interaction*, 217-222.

Bartneck C., Kanda T., Mubin O., Al Mahmud A. (2009) Does the design of a robot influence its animacy and perceived intelligence?. *Intl. J. of Social Robotics*, 1(2): 195-204.

Beck A., Stevens B., Bard K.A., Cañamero L. (2012) Emotional body language displayed by artificial agents. *ACM Trans. Interactive Intelligent Systems*, 2(1): 1–29.

Breazeal C. (2003) Emotion and sociable humanoid robots. *Intl. J. Human–Computer Studies*, 59:119–155.

Bruce A., Nourbakhsh I., Simmons R. (2002) The role of expressiveness and attention in human-robot interaction. *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, Washington DC, 4138–4142.

Bryson J.J. (2010) Robots should be slaves. In Wilks Y. (Ed.) *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues*, Natural Language Processing, 8: 63-74.

Coeckelbergh M. (2009) Personal Robots, Appearance, and Human Good: A Methodological Reflection on Roboethics*. Intl. J. of Social Robotics*, 1(3): 217-221.

Cole J. (2001) Empathy needs a face. J. of Consciousness Studies, 8(5–7): 51–68.

Cowie R. (2014) Ethical Issues in Affective Computing. *The Oxford Handbook of Affective Computing*, 334-348.

Duffy B.R. (2003) Anthropomorphism and the social robot. *Robotics and Autonomous Systems*, 42:177–190.

Embgen S., Luber M., Becker-Asano C., Ragni M., Evers V., Arras K.O. (2012) Robot-specific social cues in emotional body language. *21st IEEE Intl. Symp. on Robot and Human Interactive Communication (RO-MAN),* Paris, France, 1019-1025.

Häring M., Bee N., Andre E. (2011) Creation and evaluation of emotion expression with body movement, sound and eye color for humanoid robots. *20th IEEE Intl. Symp. on Robot and Human Interactive Communication (RO-MAN)*, Atlanta, Georgia, 204-209.

Krach S., Hegel F., Wrede B., Sagerer G., Binkofski F., Kircher T. (2008) Can machines think? Interaction and perspective taking with robots investigated via fMRI. *PloS one*, *3*(7): e2597.

Levy D. (2010) Falling in love with a companion. In Wilks Y. (Ed.) *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues*, Natural Language Processing, 8: 89-94.

Li D., Rau P.L.P., Li Y. (2010) A cross-cultural study: Effect of robot appearance and task. *Intl. J. of Social Robotics*, 2(2): 175-186.

MacDorman, K. F. & Ishiguro, H. (2006). The uncanny advantage of using androids in social and cognitive science research. Interaction Studies, 7(3), 297–337.

Moshkina L. (2012) Improving request compliance through robot affect. *26th AAAI Conf. on Artificial Intelligence (AAAI)*, 2031-2037.

National Research Council Committee to Review the Scientific Evidence on the Polygraph (2003) The Polygraph and Lie Detection. http://www.nap.edu/read/10420/chapter/1

Norman D.A. (2004) *Emotional design: Why we love (or hate) everyday things*. Basic Books, New York.

Riek L.D., Rabinowitch T.C., Chakrabarti B., Robinson P. (2009) Empathizing with robots: Fellow feeling along the anthropomorphic spectrum. *3rd IEEE Intl. Conf. on Affective Computing and Intelligent Interaction (ACII)*, 1-6.

Sabanovic S., Reeder S., Kechavarzi B. (2014) Designing robots in the wild: In situ prototype evaluation for a break management robot. *Journal of Human-Robot Interaction*, *3*(1): 70-88.

Sirkin D., Ju W. (2014) Using embodied design improvisation as a design research tool. Intl. Conf. on Human Behavior in Design, Ascona, Switzerland, 14-17.

Turkle S. (2007) Authenticity in the age of digital companions. *Interaction Studies*, *8*(3): 501-517.

Weiss A., Wurhofer D., Tscheligi M. (2009) "I love this dog"—children's emotional attachment to the robotic dog AIBO. *Intl. J. of Social Robotics*, 1(3): 243-248.

Wilks Y. (Ed.) (2010) *Close engagements with artificial companions: key social, psychological, ethical and design issues*, Natural Language Processing, 8. John Benjamins Publishing.

Złotowski J., Proudfoot D., Yogeeswaran K., Bartneck C. (2015) Anthropomorphism: opportunities and challenges in human–robot interaction. *Intl. J. of Social Robotics*, *7*(3): 347-360.

3. Robots in the Workplace

3.1. Highlights from *The Vestigial Heart*

Chapter 13, pages 85-86:

He [Leo] refused to listen to fantasies about brains pushed so hard they went mad, thrown out once they were no more than human waste, and he was even less prepared to believe them.

Now he's been able to confirm his suspicions.

[..] knowing that all the information about the prosthesis will be erased as soon as he crosses the threshold [of his workplace] makes him a bit nervous. Dr. Craft had assured him that, apart from that memory lapse, he wouldn't notice anything else. In fact, he gave him a practical demonstration when, terrified by that clause of the contract, he was about to change his mind. The timeout button, as he called it, [..] as soon as he pressed the button, it [all project knowledge] was erased not only from the screen, but also from his memory. He couldn't have explained how it worked even as a matter of life and death.

The waves of encryption the device added to the brain were innocuous, he'd checked it out. There weren't any side effects either at the time or after, so in that sense he isn't worried. What annoys him is being at the mercy of a mechanism that he doesn't understand.

Chapter 13, page 87:

[..] Lost in thought, he hasn't even realized that ROBco has taken control of the wrap-around screen that, when stretched to its maximum, covers the walls of the space, and when it asks permission to deactivate the holographic partition walls, he almost jumps out of his seat. It has the results of the comparison it was assigned this morning; not like him, who hasn't been able to stop getting side-tracked by his imminent outing. How little control, he's making a fool of himself, and this moron, with all its

neurolearning and pedigree, hasn't even learned to stop him when he's wasting time like an idiot. He looks up at the ever-watching electronic eyes of the cameras and thinks how lucky he is that they can't read his mind.

Chapter 13, page 88:

[..] he isn't in the mood right now, even if ROBco is waiting expectantly next to him, ready to act as his assistant.

"It's not worth us starting, if I have to leave in a moment." As ROBco is still staring at him insistently, he admonishes it, "I told you: I can't turn off and back on again and pick up where I left off, like you do, see if you can finally build that into your model."

"Confirmation: It was incorporated seventeen days, four hours, thirteen ..."

"Stop, stop, stop ... I've also told you several times that it's not necessary to be so precise. And if you're aware of my limitations, I don't understand why you insist on starting work."

Chapter 13, pages 92-93:

[Leo:] "I don't know, when I left they made me pass through a device that erases my memory."

[..]

[Bet:] "You see? I worry about you and you get angry. Such an intrusive protection system must be illegal. There's one at MascotER too, but it doesn't put the rights of employees at risk."

3.2. Ethical Background and Discussion

When industrial robots began to be used in production lines in factories, the main social concern was the rise of human unemployment that this could entail. Now that service robots are entering many other workplaces to perform a large variety of tasks, a similar concern has reappeared, coupled with another one: how to define the

boundaries between human and robot labor in a shared task, so that not only is throughput is maximized but, more importantly, the rights and dignity of professionals are preserved. This is exemplified by ***Leo*** struggling on two fronts: on the one hand, he fears his privacy and intellectual property rights may be violated by the mysterious timeout device installed by his employer and, on the other hand, he struggles to make ***ROBco*** `understand´ that they have different skills and, in order to optimize their collaboration, they need to do what each does best and communicate on common ground.

According to Frey and Osborne (2013)'s wide survey, about 47% of total US employment is at risk of being automated, with high wages and educational attainment exhibiting a strongly negative correlation with such risk. Based on this study, the percentage chance of being automated of some hundreds of jobs is displayed in a webpage, together with their ranking in four relevant features, namely whether the job requires negotiation, coming up with clever solutions, personally helping others, and squeezing into small spaces. Jobs found under community and social services have the lowest percentage chances of being automated, whereas telemarketing has the highest percentage. While the percentages are just approximate, the list of the 20 most risky jobs provides qualitative insight into what the future may look like. Ethical outlooks on unemployment due to robotization will be discussed under Question 3.A below.

Regarding the concern on the extent and boundaries of human-robot interaction in the workplace, we will consider three facets of it. Under Question 3.B, we will examine ways to ensure good collaboration both at the organizational level, when several professional groups are to interact with the robots, and at the level of a single individual collaborating with a robot. Then, the specific ethics issues arising in conducting research on human-robot interaction that requires experimenting with human users will be addressed under Question 3.C. Finally, the speculative issue of intellectual property rights on inventions derived from the joint work of robots and humans will be briefly presented under Question 3.D.

Question 3.A – Would robots primarily create or destroy jobs?

Concern about job loss is not specific to robotics; it appears whenever the use of machines displaces human labor, and it can be traced back to the agricultural and industrial revolutions and, more recently, to the Internet revolution. The standard response is that human workers are thus freed from dangerous, dirty, or dull tasks (the infamous three D's) to be able to undertake "higher value" jobs, mostly in the design, programming, deployment, maintenance and use of these new technologies. Gorle and Clive (2011) provide evidence of the direct and indirect creation of employment due to the use of robots in several sectors. However, this positive trend has a downside: the technological divide. Most of the displaced workers won't be able to take on the new jobs. In developed countries, the skill shift may take at least one generation and, for underdeveloped societies, the economic gap may become insurmountable. Therefore, at this macro-political scale, the progressive robotization of labor should mandatorily be accompanied by appropriate social measures (e.g., a more equitable distribution of work and resources) that balance out its impact, especially for the most disadvantaged.

At the enterprise level, it is important to promote an ethical business culture in the robotized workplace. Chijindu and Inyiama (2012) list six measures that could be adopted by governments, industries and unions, with particular emphasis on emerging economies. For instance, engineering trade unions could make sure that corporations have proper plans to re-train employees for the 'higher value' jobs that are said to emerge. Likewise, the way to introduce robots in an organization where they would interact with several groups of professionals needs to be well thought out in advance so as to avoid undesirable preventable effects, as described under the following question.

Question 3.B – How should work be organized to optimize human-robot collaboration?

Dyadic human-robot interaction in a working environment has some particularities that need to be carefully analyzed. We will distinguish between human-robot

collaboration and experimentation, which will be dealt under the next question. Concerning the former, it is known from human studies that good collaborations occur when people share the same goals but have unique roles, when they can learn how to communicate effectively about the problem and solution spaces, when they come to respect and trust their collaborator's responses, and when they begin to enjoy spending time working with their partners. This has been termed "relationship potential" by Bernstein *et al.* (2007). A successful human-robot relationship will require the human to develop a set of beliefs about the robot that aid collaboration and will require the robot to clearly communicate capabilities relevant to the collaboration.

Thus, it is important that employees see robots as tools, not replacements for their jobs, and develop a sense of "ownership" that favors commitment to take out the best of human-robot collaboration. To this aim, technology impact (acceptance or resistance) in the workplace should be addressed already at design time [Borenstein 2010; Salvini *et al.* 2010], and aspects such as whether robots will improve the workers' quality of life and working conditions must be taken into consideration.

In this regard, Decker (2007) states that moral reasons should take priority over considerations of utility, so that the human actor in human-robot cooperation should never be instrumentalized, i.e., used solely as a means to achieve a particular end. This follows from Kant's principle: «Act in such a way that you treat humanity, whether in your own person or in the person of another, always as an end and never as a means.»

Beyond dyadic interactions, Barrett *et al.* (2012) analyze how introducing a pharmaceutical dispensing robot in a hospital changed the boundary dynamics of three occupational groups (pharmacists, technicians and assistants) and pay attention not only to the most dramatic aspects of conflict and resistance, but also to more subtle ones like some groups expanding/shrinking their jurisdiction, expertise and professional standing. These authors advocate for taking into account this holistic view when planning the introduction of robots in the workplace so as to organize work efficiently without inadvertently impairing the identity, status and authority of any of the involved occupational groups.

Also in a hospital context, Mutlu and Forlizzi (2008) found diametrically different responses in two departments to robots taking out laundry from patients' rooms. In the postpartum department housekeepers happily stopped to load a robot whenever it arrived, whereas in the cancer unit, low tolerance for interruptions, a discrepancy between the perceived cost and benefits of using the robot, and breakdowns due to high traffic and clutter in the robot's path caused the robot to have a negative impact on the workflow and staff resistance. Dietsch (2010) concludes that robots should conform to the atmosphere and protocols of the workplace and puts forward a common sense formula: «Robots that wait patiently for people's schedules are affirming. Robots that demand people meet their schedules are not.»

Bahn *et al.* (2015) discuss an experiment in which a humanoid robot was placed to work alongside humans in an assembly line. The unusual aspect is that the robot served as a quality inspector and provided feedback about how well the human workers performed their work. Feedback took four forms: positive, negative, neutral and contradictory, whereby facial expression and gestures contradicted the textual report. In general, the conclusions corroborated what one would expect, except perhaps that negative form correlated with attributing intelligence to the robot and contradictory feedback was interpreted as mockery. Anyhow, very few subjects participated in the experiment, thus the results are very preliminary, but they suggest ways to orient research on ethic social human-robot interaction in the workplace.

Question 3.C – Do experiments on human-robot interaction (HRI) require specific oversight?

As the last paragraph unveils, research on interactive robots raises a new range of ethical concerns in relation to humans, beyond those typically considered in industrial robotics. These are mainly of social and psychological nature, and some are shared with experimentation on human-computer interaction (HCI). In this field, Punchoojit and Hongwarittorrn (2015) identified thirteen categories of concerns, among which we highlight that research trials need to be approved by established ethics committees, participants should be thoroughly informed and their self-determination ensured by signing consent forms, individual differences (cultural, age-related, disabilities, etc.)

should be taken into account in the design of the experiments, and privacy of their data must be guaranteed through anonymization or other data protection procedures.

Concerns specific to experimentation on HRI are minimizing risks of physical damage, foreseeing possible emotional reactions of participants towards robots and how to handle them, and avoiding deceit especially in the case of vulnerable groups such as children, elderly or disabled people.

Deceit may come from the use of Wizard-of-Oz, a technique frequently employed by HRI practitioners, whereby a person remotely operates a robot puppeteering many of its attributes (speech, nonverbal behavior, navigation, manipulation, etc.) in order to collect experimental data on attitudes towards robots. Riek (2012) provides detailed guidelines to ensure careful use of Wizard-of-Oz in HRI research and remarks that the possible fostering of inappropriate expectations among users must be taken into account, in the same way as with humanoid morphology and functionality, as discussed in Section 1. The Principles of robotics, issued by the Engineering and Physical Sciences Research Council of the UK, state that the best way to protect participants in experiments is to «guarantee a way for them to 'lift the curtain' (to use the metaphor from The Wizard of Oz)».

In sum, provisions should be made to ensure that participants in HRI research experiments are being treated with respect and integrity, and are having their rights protected; as abusing the participants can bring disrepute to the whole research community.

Question 3.D – Do intellectual property laws need to be adapted for human-robot collaborations?

This is a far-fetched speculative issue.

When a robot is equipped with learning capabilities, its behavior may become unpredictable to some extent, especially in the long run due to the effects of the experiences gathered. Responsibility for its actions would then be split between the programmer, the user, and possibly others: for example, those that had previously

trained it. In Chapter 18 of the novel, ***ROBco*** comes up with some hints that help ***Leo*** to find a solution to a problem. In this case, it was ***Leo*** himself who trained the robot. Anyhow, the possibility that future robots show creative thinking as a result of the interplay of their learning algorithms with their experiences should not be dismissed.

As Bernstein *et al.* (2007) mention, if we are going to treat humans and robots as legitimate collaborators, they deserve to be evaluated as a collaborative unit. Thus, the responsibility for their joint successes and pitfalls will be shared, and in trying to improve their performance, credit would need to be assigned to one or the other.

In a currently more realistic scenario, Hinds *et al.* (2004) studied under which conditions people relied on and ceded responsibility to a robot coworker. Not surprisingly, the results show that participants retained more responsibility when working with a machine-like as compared with a humanoid robot, leading the authors to suggest that in settings where people have to share or delegate responsibility and when complacency is not a major concern, humanoid robots may be more appropriate. This is a debatable conclusion that may incur in contradiction with the need to avoid deceit in interacting with a robot.

All in all, as Patel (2016) concludes: «The best solutions are always going to come from minds and machines working together.» The challenge is, of course, not to fall into complete technological dependency.

3.3. Revisiting issues

The readings from pages 87 and 88 in Chapter 13 of *The Vestigial Heart* touch on two issues discussed in Section 1, namely nudging and robot adaptation to the user, respectively. The advantages and risks of nudging were discussed under Question 1.D, and here ***Leo*** would like ***ROBco*** «to stop him when he's wasting time like an idiot.» In relation to the need of fast robot adaptation to the user, ***Leo*** complains that ***ROBco*** has not yet built into its user model that he cannot switch from one task to another so quickly, and that he doesn't need so much precision in the robot explanations.

References

Banh, A., Rea, D. J., Young, J. E., & Sharlin, E. (2015). Inspector Baxter: The Social Aspects of Integrating a Robot as a Quality Inspector in an Assembly Line. In Proceedings of the 3rd International Conference on Human-Agent Interaction, pp. 19-26.

Barrett M., Oborn E., Orlikowski W.J., Yates J. (2012) Reconfiguring boundary relations: Robotic innovations in pharmacy work. *Organization Science*, 23(5), 1448-1466.

Bernstein D., Crowley K., Nourbakhsh I. (2007) Working with a robot: Exploring relationship potential in human–robot systems. *Interaction Studies*, 8(3), 465-482.

Borenstein J. (2010) Computing Ethics. Work Life in the Robotic Age. *Communications of the ACM*, 53(7): 30-31.

Chijindu V.C., Inyiama H.C. (2012) Social implications of robots: An overview, *International Journal of Physical Sciences*, 7(8): 1270-1275.

Decker M. (2007) Can humans be replaced by autonomous robots? Ethical reflections in the framework of an interdisciplinary technology assessment. *Workshop on Roboethics*, IEEE Intl. Conf. on Robotics and Automation (ICRA).

Dietsch J. (2010) People meeting robots in the workplace [industrial activities]. *IEEE Robotics & Automation Magazine*, 17(2), 15-16.

Frey C.B., Osborne M.A. (2013) The future of employment: how susceptible are jobs to computerisation. *Workshop on "Machines and Employment"*, Oxford Martin Programme on the Impacts of Future Technology, Oxford Martin School, University of Oxford.

Gorle, P. and Clive, A. (2011). Positive impact of industrial robots on employment. Metra Martech. www.ifr.org/uploads/media/Metra_Martech_Study_on_robots_02.pdf

Hinds P.J., Roberts T.L., Jones H. (2004) Whose job is it anyway? A study of human-robot interaction in a collaborative task. *Human-Computer Interaction*, 19(1), 151-181.

Mutlu B., Forlizzi J. (2008) Robots in organizations: the role of workflow, social, and environmental factors in human-robot interaction. *3rd ACM/IEEE Intl. Conf. on Human-Robot Interaction (HRI)*, pp. 287-294.

Patel P. (2016) What AI Experts Say Smart Machines Will Do to Human Jobs, *IEEE Spectrum*.

Punchoojit L., Hongwarittorrn N. (2015) Research ethics in human-computer interaction: A review of ethical concerns in the past five years. *2nd National Foundation for Science and Technology Development Conference on Information and Computer Science (NICS)*, pp. 180-185.

Riek L.D. (2012) Wizard of Oz studies in HRI: A systematic review and new reporting guidelines. *J. of Human Robot Interaction*, 1(1): 119-136.

Salvini P., Laschi C., Dario P. (2010) Design for acceptability: Improving robots' coexistence in human society. *Intl. J. of Social Robotics*, 2(4): 451–460.

4. Robots in Education

4.1. Highlights from *The Vestigial Heart*

Chapter 14, pages 96-97:

... she can connect with the monitoring circuit [at school] reserved for parents and observe what she [Celia] is doing right now. Silvana hesitates for an instant. She should decline the offer without a second thought: she's denounced technology that violates privacy so many times, for how it undermines the privacy of the weakest among us; if someone from the ComU were to catch her spying, she would die of shame.

[..]

"Can we hear what they're saying?"

[Lu:] "No, no. It would violate the rights of the child, you should know that," she says, giving her a suspicious look.

"Sorry, but I don't understand: the images are public but the sounds are private?"

"Come on! Who said anything about it being public?" She seems to be outraged, it must be a hot-button issue. "Parents have a right to check on the physical integrity of their children at any time, that's all. If someone breaks that rule, with lip-reading programs or any other tricks, their connection to the circuit is cut off forever."

Chapter 14, page 98:

[Silvana:] "What are those figures they keep bumping into, they look like mannequins."

[Lu:] "They're for practicing socialization." She stops for a moment as if she can't be bothered to explain it. "They stage a situation and the kids have to practice until they learn to behave properly automatically. It's one of the most innovative

activities in the school, they call it social-conduct training; they recommended it for Celia and it's been good for her, in just a few days she's caught up with the rest."

It occurs to Silvana that it's like learning to drive, only that instead of controlling a machine that navigates among other machines, it's navigation among people that is automated ...

Chapter 14, pages 99-100:

... the teacher has labeled Celia a rebel because, ignoring his advice, she insists on competing with machines.

[..] It was the EDUsys that started sending out alarm signals because she didn't use it as she was supposed to. Apparently she hasn't taken to the net's search mechanisms and, faced with a question, she stops and thinks about it, trying to make up an answer, instead of trusting what other people have thought before. "Imagine if we all had to start from scratch!" the teacher exclaimed, annoyed. He himself doesn't have most of the knowledge they are working on, he told her with pride, that's what EDUsys is for ...

[..] Silvana will have to work hand in hand with ROBbie, because everyone knows that if the child turns out to be a rule-breaker, the robot must learn to restrain them, to counteract their impulses, put them on the right track ... Robots are customizable for a reason, they have to complement their PROPs to make a good team.

That's the last thing Silvana expected to hear, that she'll have to train a robot!

Chapter 16, pages 112-113:

[..] They've hired a home tutor for me. I can tell you're surprised and I can guess what you're thinking: "For you? But you've always done so well in school." But, you know what? There are no subjects nowadays, they just teach you to use EDUsys and to behave. You don't have to memorize anything, like before in geography and history, and you don't have to learn formulas either, like we did in math, the ROBs do that. It's like they're teaching us to play, first on our own with the computer and then in a group in the socialization room.

[..]

According to her [Silvana], my problem comes from the fact that I react differently than kids that are around these days, and that's why EDUsys has problems programming my education. I was really pleased she said it was the robot that had problems, not me.

Chapter 17, page 121:

[Silvana:] "There must be some company behind it," she announces severely. "Which one is it?"

[Leo:] "Okay"—hiding information can only be counterproductive—"CraftER."

"I knew it! The ones who make more and more intelligent robots for ever-stupider humans … and now you want to destroy our creativity by passing it on to robots?"

"No, no, it's quite the opposite. It's about strengthening human creativity, making a kind of devil's advocate that spurs it on."

"You lot are always nit-picking: you're not trying to replace anyone, just broaden their abilities, that's why you rush to use euphemisms, like assistant or helper, instead of saying executor or usurper, which is what they'll end up being."

Chapter 22, pages 155 and 160-161:

[Xis:] "… I'll just tell ROBix we're leaving and then we can go."

[Celia:] "No, I told you already"—many more moments like this and she'll regret having persuaded her—"our ROBs can't know anything about this, they might stop us or tell on us."

"My ROBix … never!"

"How do you know? Have you ever tried it? Anyway, we already agreed, we're not going to tell them anything."

"You're right, but …"—she's all stressed out—"it always has to know where I am."

"Why?"

"Because … I don't know, how else will it keep an eye on me?"

[..]

"I'm going to fall, I'm going to fall!" Xis' hands grip the pole that separates them from the abyss, turning her knuckles white.

"Calm down, Xis, don't look." She gently takes hold of her from behind and makes her turn around. "It's not that different than what you see every day from the aero'bus that takes us to school."

"But there's no protection here! I can see there's nothing, nothing"—she whimpers, terrified, stretching her arms out as if they too were part of the horror. "I want ROBix right now. I want to go back."

[..]

[Xis:] "Open it, open it! I want to get out." She strikes the membrane with both hands, beside herself. "I want ROBix. It'll know what to do. I need to talk to it …"

"Calm down. We're not in any danger. We'll just wait here quietly until another aero'car parks and we'll get out. We could be out really soon."

"We'll never get out! You're from another century, you have no idea how anything works."

4.2. Ethical Background and Discussion

Telepresence or semi-autonomous robots to teach music or foreign languages are regarded as useful aids in the classroom [Kanda *et al.* 2004; Chang *et al.* 2010], as are educational robots for initiating young children into programming or for enforcing teamwork to consolidate concepts from various disciplines [Mitnik *et al.* 2008]. Polemics arises when autonomous robotic assistants are envisaged to take the role of human teachers in the transmission of cultural values and critical thinking. How could a machine motivate students or provide personal moral example without the

experiences of life? How will children learn to empathize and to reason, not just logically, but also emotionally?

The influence of robots in education goes well beyond the classroom, and spans from robot nannies to robot companions for teenagers and adults. As the epigraph in the novel reads, «It is the relationships that we have constructed which in turn shape us». The philosopher Robert C. Solomon (*The Passions*, 1977) was referring to human relationships, namely with relatives, friends and teachers, but in the context of human-robot interaction the quotation acquires another meaning: the robot teachers and companions that we are constructing will in turn shape us and future generations.

New educational technologies will help children develop new capacities, possibly to the detriment of others. Hence, a gradual evolution of human thought, feelings and relationships will naturally and ineludibly take place. In this regard, Sharkey and Sharkey (2010) examine the ethical concerns raised by robots acting as surrogate carers for children, focusing on the consequences for their psychological and emotional wellbeing. Since, as these authors say, it would be unethical to conduct experiments on long-term care of children by robots, they turn to developmental psychology to elucidate what a child needs for a successful relationship with a carer, centering their attention on the pathology and causes of attachment disorders. This issue will be addressed under Question 4.A below.

A note of caution is in place here: the inappropriateness of conducting the long-term experiments above should not open the door to such experimentation taking place uncontrolled and online in the homes of many incautious consumers, due to the pressure of the market. Sharkey and Sharkey (2010) mention a long chain of responsibility and it is worth emphasizing that, since education and feelings are at the core of human nature, the type of robot carers to devise should not be a specialized debate confined to designers and producers, but one that should concern everyone. Consumers should be able to distinguish robots stimulating the child's best abilities from those just creating dependence, spoiling children or becoming a substitute for parenting.

Question 4.A – Are there limits to what a robot can teach?

Largely simplifying, we can distinguish three types of competences children need to acquire: cognitive, social, and emotional. Robots can help teach some of the former ones, maybe partly the second, but not the latter ones. On the positive side, Mitnik *et al.* (2008) report an experiment in which «the robot was not only able to guide the team of students to pursue a common goal, but it was also able to provide unique capabilities to each student, fostering collaboration and inhibiting free-riding behaviors». Individualized assistance, long and detailed learning traces, and nudging to favor socialization are three useful features robot teachers can provide. Examples of such beneficial nudging to favor socialization will be provided under Question 4.C.

What are the risks? At the cognitive level, children may not learn to acquire knowledge and reason about it, but rely on the robot's large memory. Turkle (2010) reports that Howard, fourteen, said a robot would be better able to grasp the intricacies in the day of a high-school student than his father, alleging «its database would be larger than Dad's. Dad has knowledge of basic things, but not enough of high school». Chapters 14 and 16 in the novel illustrate the limitations of education plans centered on machines and personalized to each student by machines, leading ***Celia*** to conclude that ***EDUsys*** has problems programming her education because she reacts differently than kids that are around those days: faced with a question, she stops and thinks about it, trying to make up an answer, instead of searching and trusting the machine's output.

In regards to social skills and emotions, Sharkey and Sharkey (2010) fear that inappropriate and exclusive care of a child by a robot could lead to behavior indicative of Reactive Attachment Disorder (RAD). Such disorder prevents appropriate social relatedness, as manifest either in (i) failure to appropriately initiate or respond to social encounters, or (2) indiscriminate sociability or diffuse attachment. Other authors express a similar concern that children interacting too much with machines at early stages and deprived of human care will not develop empathy towards others and won't be able to interpret other people's feelings. At ***Celia***'s school, students are subject to an extreme, mechanical form of socialization training, which seeks to trigger a proper automatic reaction in front of people, thus emptying the encounter of any

feelings or deep connection. Having been subject to this training, it is not surprising that ***Xis*** shows signs of suffering from RAD.

Along the same line, Veruggio and Operto (2008) alert that robot toys could cause psychological problems in children, such as loss of touch with the real world, confusion between natural and artificial, and confusion between real and imaginary. Thus, designers of such robots should take into account the kind of interactions that are appropriate for a child to engage in at certain ages or stages of development.

Question 4.B – Where is the boundary between helping and creating dependency?

The key education dilemma between protecting and promoting autonomy in children appears also in the context of child-robot interactions. Sharkey and Sharkey (2010) ask «if a child was about to run across the road into heavy oncoming traffic and a robot could stop her, should it not do so?» Of course, this is an extreme case of "protection", but many other situations can be envisaged in which risks need to be taken for kids to acquire a sense of danger and be able to learn to take care of themselves. The complete dependency of ***Xis*** on ***ROBix*** shows clearly in Chapter 22, where she admits that she needs to be watched out all the time by her robot, which always knows what to do and will free her of all potential dangers.

Some artificial companions for children, such as the Junior Companion mentioned by Wilks (2010), sound like Big Brother watching you, raising not only privacy issues, but also triggering the question of how would children feel if their parents knew what they were doing all the time? In Chapter 14, ***Lu*** takes for granted that parents have the right to constantly monitor what their children are doing, which may prevent them from learning to behave autonomously and impair their decision-making abilities. She further encourages child dependence by telling ***Silvana*** that she should teach ***Celia*** and ***ROBbie*** as a team, so that the robot learns to cover up the flaws of the girl.

It is worth mentioning that dependency may appear in both directions in the caring relation, namely the carer may also suffer from it. Citing Turkle (2007), «tamagotchis demonstrated a fundamental truth of human-machine psychology. When it comes to

bonding with computers, nurturance (an application that can eliminate its competitors)». Robot designers must balance the risk of creating dependency with the beneficial effect that caring as symbolic play may have for child development, without disregarding the fact that teaching a robot is an effective way for a child to learn [Tanaka and Kimura 2009].

Pearson and Borenstein (2014) examine the ways in which particular design features (e.g., gendered appearance, humanlike behavior, etc.) may affect children's short- and long-term development, so as to orient design decisions to promote their physical, psychological, and emotional health. They question whether concerns about the uncanny valley hypothesis (discussed in Section 2) are still relevant today, primarily in the context of designing robots for children.

Question 4.C – Who should define the values robot teachers would transmit and encourage?

Robot nudging behavior as a possible way to control and manipulate people was discussed under Question 1.D. Here we focus on the values that can be taught to children in this way, and raise the issue of whether robots should come with some predefined values encoded and to what extent parents and teachers should have the right to modify such encoding.

As mentioned earlier, ***Celia***'s classmates are taught to behave properly in an automatic manner. Of course, better ways of teaching valuable social behavior can be imagined. For instance, a robot could smile or display other cues that encourage the sharing of toys between playmates, and mimic expressions of disappointment whenever a child refuses to share. These are mild forms of promoting generosity and altruism at early stages in development. Likewise, robots could nudge children to interact with other children with whom they don't associate so as to avoid forming cliques. This will discourage discrimination and unequal treatment in the playground. May parents/teachers regulate the degree of reward and punishment that such robot nudging entails? The extent to which robot moral behavior should be tunable will be discussed later under Question 6.B.

To illustrate how robots could promote social justice, Borenstein and Arkin (2016) use the two examples above: toy sharing and clique avoidance. These researchers claim that robots could nurture *inequality aversion* in children (a feeling developed between the ages of 3 and 8) by reinforcing proper social norms and etiquette during playtime. Furthermore, the robot could nudge a child to interact with other children with whom he/she is not as used to engaging in an effort to avoid *parochialism*, i.e., favoritism towards the child's own social group. As these authors note, adults are not always successful at displaying these good pro-social attitudes, in part because they can have difficulty suppressing their anger or frustration. Thus, a potential advantage of robots assisting in this effort is that they will not display negative emotions. Nothing prevents them from leading by example.

Besides social values, another human trait that is highly valued nowadays is creativity. However, some voices alert of the increasing risk that technology may prevail over creativity in human development. This is what presumably has happened in the world in which ***Celia*** wakes up, where this human trait is almost extinguished, this being the reason why her creativity strongly catches the attention of both ***Leo*** and ***Silvana***. This important theme, underlying the entire novel, is made particularly explicit in the words of ***Silvana*** in Chapter 17, as well as in ***Celia***'s performance when subject to ***Leo***'s tests in Chapter 19.

Question 4.D – What should the relationship be between robot teachers and human teachers?

In their review of the field of robots in education, Mubin *et al.* (2013) distinguish three roles the robot may take in the learning activity, namely tool, peer, or tutor. The role of the teacher is different in each case: if the robot is used as a tool (e.g., to teach programming or sensor physics), the teacher takes on the role of a facilitator, whereas if the robot acts as a peer (essentially providing encouragement when the student performs correctly), then the onus is on the teacher to transfer knowledge. The robot role that raises more concerns is that of tutor, which the authors view as an assistant that adapts exercises to each student or helps remember vocabulary, for example. Their message is that robots are not intended to replace human teachers but can bring

an added value to the classroom in the form of a stimulating, engaging and instructive teaching aid. To this end, it is urgent to develop appropriate curricula and materials for training teaching staff to share activities in the classroom with a robot.

Tanaka *et al.* (2007) focus on the use of robots as a tool and, after immersing a social robot in a classroom of toddlers for more than 5 months, they conclude that robot technology has great potential in educational settings assisting teachers and enriching the classroom environment. These authors stress that robots should be designed to assist and support teachers' educational activities together with them and under the control of them.

There is no doubt that robots can have a large repertoire of exercises, both intellectual and physical, and rehearse them with infinite patience, thus they can be very valuable in handling simple tasks that take up teachers' precious classroom time. Furthermore, they can be very handy at building a model of each student, and keeping long traces of their progress and attitude. Now the question arises of whether a robotic teaching assistant should team up with the teacher or with the students. On the one hand, having access to all the student information resulting from their interaction with the robot can be very useful for human teachers to provide personalized assistance. But, on the other hand, in order to be trusted by children, robotic assistants must not disclose their "secrets" to the teacher. Establishing the balance between the former and the latter is a difficult task.

This confidentiality issue is discussed by Sharkey and Sharkey (2010) and illustrated in Chapter 22 when ***Xis*** says that ***ROBix*** would never tell on her, and ***Celia*** questions this statement alleging that she never tried. Moreover, in Chapter 14, when ***Lu*** and ***Silvana*** are watching what ***Celia*** is doing at school, they make clear that conflicts of interest regarding privacy have been the object of careful regulation by law in the described futuristic society.

4.3. Revisiting issues

The highlighted paragraphs in Chapter 17 permit revisiting some of the issues discussed in Section 1; in particular, potential downsides in the design of robot assistants. On the contrary, the relation of ***Leo*** with his assistant ***ROBco*** in Chapter 18 illustrates how a robot could enhance the capacities of its user (as already mentioned under Question 3.D), although in this case under the control of ***Dr. Craft***, thus also touching upon the discussion under Question 1.D. Finally, the discovery in Chapter 21 that ***Leo*** has been used as guinea pig by ***Dr. Craft*** stresses the concerns previously considered under Question 3.C.

References

Borenstein J., Arkin R. (2016) Robotic Nudges: The Ethics of Engineering a More Socially Just Human Being. *Science and engineering ethics*, 22(1): 31-46.

Chang C.-W., Lee J.-H., Chao P.-Y., Wang C.-Y., Chen G.-D. (2010) Exploring the Possibility of Using Humanoid Robots as Instructional Tools for Teaching a Second Language in Primary School. *Educational Technology & Society*, 13 (2), 13–24.

Kanda T., Hirano T., Eaton D., Ishiguro H. (2004) Interactive robots as social partners and peer tutors for children: A field trial. *Human-computer interaction*, 19(1): 61-84.

Mitnik R., Nussbaum M., Soto A. (2008) An autonomous educational mobile robot mediator. *Autonomous Robots*, 25(4): 367-382.

Mubin O., Stevens C.J., Shahid S., Al Mahmud A., Dong J.J. (2013) A review of the applicability of robots in education. *Journal of Technology in Education and Learning*, 1: 209-0015.

Pearson Y., Borenstein J. (2014) Creating "companions" for children: the ethics of designing esthetic features for robots. *AI & society*, 29(1): 23-31.

Sharkey, N. & Sharkey, A. (2010) The crying shame of robot nannies: An ethical appraisal. *Interaction Studies*, 11(2): 161-190.

Tanaka F., Cicourel A., Movellan J.R. (2007) Socialization between toddlers and robots at an early childhood education center. *Proceedings of the National Academy of Sciences*, 104(46): 17954-17958.

Tanaka F., Kimura T. (2009) The Use of Robots in Early Education: A Scenario Based on Ethical Consideration. *Proc. 18th IEEE Intl. Symp. on Robot and Human Interactive Communication (ROMAN 2009)*, Toyama, Japan, pp. 558-560.

Turkle S. (2007) Authenticity in the Age of Digital Companions. *Interaction Studies*, 8(3): 501-51.

Turkle S. (2010) In good company? On the threshold of robotic companions. In Wilks Y. (ed.): *Close engagements with artificial companions: key social, psychological, ethical and design issues*, pp. 3-10, Amsterdam, The Netherlands: John Benjamins Publishing Company.

Veruggio G., Operto F. (2008) Roboethics: Social and Ethical Implications of Robotics. In: Siciliano B., Khatib O. (eds) *Handbook of Robotics*, pp. 1499-1524, Springer.

Wilks Y. (2010) Introducing Artificial Companions. In Wilks Y. (ed.): *Close engagements with artificial companions: key social, psychological, ethical and design issues*. Amsterdam, The Netherlands: John Benjamins Publishing Company.

5. Human-robot interaction and human dignity

5.1. Highlights from *The Vestigial Heart*

Chapter 25, pages 177-178:

[Leo:] "If you'll allow me to ask a question … Why is it so hard for you people to accept that machines can perform some tasks better than we can?"

[..]

[Silvana:] … And, by the way, who are 'we'?"

[..] "The anti-techno view, I mean." He pauses, as if he doesn't dare give voice to the thought that's playing in his head. "You know … you're the first one I've met in person."

[..] "And what? Do I seem very eccentric to you?" She turns to face him with such force that she hurts herself on the seat's ergonomic springs. This cutting-edge comfort technology is all very well, but the designer never even anticipated that passengers might want to talk to each other.

Chapter 25, pages 179-181:

[Silvana:] "How can you be so frivolous? You don't even know if working for that company might be detrimental to you. What are you trying to do now? Making robots with feelings … and you have to suck them out of a little girl?"

[Leo:] "No, no, please. I tried to explain it to you the day of the get-together: it's about boosting human creativity"—he hopes saying it like that will make it sound better—"by giving people an assistant that stimulates them."

"Very nice … but do you believe in it?"

Suddenly a neutral voice interrupts the conversation:

"Information: I am an example."

The shock makes Silvana jump and stab herself on the springs again. Absorbed in the conversation, she'd forgotten they had a silent witness.

"What is it saying? That what you took out of Celia you put into this thing?" she shouts, pointing at the robot, one wrong move away from destroying her back in the process.

[..] "Calm down, please. Nobody's hurt Celia. What ROBco means is that it has been fitted with a prototype of the prosthesis we're developing. You see, I'm the guinea pig, not the girl," he concludes, with resignation.

Poor naive boy, Silvana thinks, it's quite possible that is the case.

"And you're okay with that ... ?"

"Yes, think about it: it's a device designed by me that helps expand my capabilities. What more could I want?" He never would have thought he'd end up defending the Doctor's project so convincingly.

"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 that ends up hogging people's most intimate time and space. You said it yourself: you don't need anything else ... and, in the end, you become wooden like them."

"That's what really gets me about the anti-techno lot"—Leo can't take this anymore—"you confuse everything, you get it all mixed up. First off, I was talking about expanding capabilities, not augmenting them. The machines you're so fond of are useful, sure, but they only magnify what we already have. I'm talking about creating new skills, broadening the range of what we can do. For example ROBco ..."

*[..] "*Question*: Would you like a suggestion?" Upon receiving Leo's assent, it goes on. "Try not to repeat yourself. I have already been used as an example and it is*

obvious that she does not like ROBs. Look for another example, one that appeals to her more."

"Don't you find it degrading when it talks to you like that?"

"Why? It's given me some good advice. Quite the opposite, I'm pleased the prosthesis is working."

Without a doubt this idiot is as wooden inside as he is on the outside. Now he'll make an effort to obey the robot.

Chapter 28, pages 204-205:

[Leo:] "What do you take me for? That's the ROB leaving, not me."

[Silvana:] "Of course, I forgot, you built it, so you've already mastered everything it knows how to do."

"Not quite. He accumulates knowledge from lots of different people."

"Okay, okay, I meant that you're not a typical PROP, you take the initiative, not the other way around, like usual."

"I don't understand. All ROBs serve people."

"Exactly. It's just that the service is often poisoned. Why do you think we're against those mechanical contraptions?" She feels she can say this now that the dummy's not around. "Because we're snobs? Well, no." She's set her course and there's no stopping her now. "Overprotective robots produce spoiled people, slaves produce 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 sell."

5.2. Ethical Background and Discussion

Users would expect robot assistants to have the basic interaction competencies to deal with ethically-sensitive situations. This is especially critical in the case of robot caregivers for vulnerable groups, such as children, mentally disabled or elderly people. For example, the advantage of robots being always available and very reliable to

provide physical support in patient transfer operations must be balanced with the need to avoid eliciting feelings of objectification and loss of control; thus, robots should not lift or move people around without consulting them. Physical contact with robots appears as a sensitive issue when ***ROBco*** places a helmet on ***Celia***'s head in Chapter 19.

Sharkey and Sharkey (2014) identified six major issues to be considered before deploying robot technology in eldercare: (i) opportunities for human social contact could be reduced, and elderly people could be more neglected by society and their families than before; (ii) risk of objectification, as we mentioned above; (iii) loss of privacy; (iv) restriction of personal liberty; (v) deception and infantilization that might result from encouraging interaction with robots as if they were companions; and (vi) attribution of responsibility if things went wrong, which opens up the key general concern about the limits of robot decision making in relation to the user's state of mind as addressed under Question 5.A.

Note that most of these issues are not specific of robots for eldercare, and apply as well to robot companions and even more generally to other types of human-machine interaction… or non-interaction through automatic decision making. This brings us to smart city technologies, such as ambient intelligence and the internet of things, which make it possible for robots to share databases, procedures and experiences, i.e., maps of visited buildings, object models and instructions of use for all kinds of commercial products, manipulation skills, and other acquired information and expertise, which can be very handy in some cases but that, leaving the human out of the control loop, may restrain the freedom and privacy of citizens.

Question 5.A – Could robot decision-making undermine human freedom and dignity?

A feeling of vulnerability similar to that caused by an unforeseen physical contact with a robot may occur at the cognitive level, the solution in this case being much more involved than simply informing the user. Not only is the complexity of the information to be transmitted much higher, but, more importantly, the extent to which a robot

should decide and convey its decisions to users depends on their state of mind, which is difficult to evaluate and evolves over time.

Even in the restricted domain of automatic emotion detection—a technology not yet well developed—errors in the interpretation of human mood expressions could strongly impair communication with the user and, more severely, entail danger for the person (e.g., failing to call an emergency service). As Cowie (2015) mentions, the problem is not new, a classical example involving 'lie detectors': despite widespread belief in their powers, they were actually much more likely to stigmatize the innocent than to pinpoint the guilty.

Thus, procedures must be devised to ensure that users are not subjected to actions they do not deserve, or not receive responses that they ought to. On a milder scale, provisions should be made for robots to always use respectful language and never intimidate users. In the last highlights above taken from Chapter 25, ***Silvana*** reacts to what she feels is a harsh piece of advice from ***ROBco*** by asking ***Leo*** if he doesn't find it degrading that the robot talks to him like that.

Principles of robotics, issued by the Engineering and Physical Sciences Research Council of the UK, as already mentioned in Section 1, state «a robot used in the care of a vulnerable individual may well be usefully designed to collect information about that person 24/7 and transmit it to hospitals for medical purposes. But the benefit of this must be balanced against that person's right to privacy and to control their own life e.g. refusing treatment.»

A related issue where balance is also needed appeared when discussing Question 1.D, namely whether it is ethically admissible to design robots that can influence human behavior, and if so, whether users must always be aware of robot nudging and how much control they should have over it.

In summary, there is wide consensus that robots and computational systems should be designed in ways that (i) do not denigrate the user to machine-like status, and (ii) do not impersonate human agency by attempting to mimic intentional states leading to deception [Lichocki *et al.* 2011]. Moreover, people should be able to decide whether

they wish to interact with these artificial "creatures" and, in case they decide they want to interact only with humans, they should be given the freedom to do so, a guideline that is not easy to implement, as the many companies using chatbots to provide customer support demonstrate.

Question 5.B – Is it acceptable for robots to behave as emotional surrogates? If so, in what cases?

The idea of robot companionship seems natural to some people and almost obscene to others. Levy (2007), in his provocative book and a review of the state of affairs ten years later [Cheok *et al.* 2017], maintains that many people will no doubt fall in love with robots and that this is completely normal. On the other hand, Bryson (2010) argues that artificial companions should just be servants, machines that you should be able to switch off whenever you like. Sullins (2012) holds an intermediate position in that he accepts people will relate to love machines, and he proposes some ethic design principles to limit the manipulation of human psychology when it comes to building sex robots and simulating love in such machines.

Given the sometimes painful and capricious nature of human relationships, it is not surprising that some might prefer to share their life with a robot, which would have predictable behavior and never criticize, cheat, or disclose their intimacy. This may be acceptable for an adult in full command of their mental faculties, but emotional surrogates should generally be avoided in the case of vulnerable users, and especially children.

The illusion of emotions may have undesired effects on people that are psychologically weak, immature, diminished, or with no technological background, and the risk that they end up being manipulated must be minimized [Boden *et al.* 2017]. Turkle (2007) advises never to disregard that, although the machine may only have simulated emotion, the feelings it elicits are real. Like in other ethical issues discussed up to now, a balance needs to be reached here since, for instance, human caregivers sometimes simulate affection to improve their patient's well-being, and thus robots may also be allowed to do so under similar circumstances.

Let's stress that there is a difference between simulating affection and showing emotional intelligence. The latter entails capturing the emotional state of the user and acting accordingly, which can be very handy in some healthcare situations, but dangerous in the case of interpretation errors as discussed under the preceding Question 5.A.

Robot companionship, even for people with full adult judgment, may have some social consequences as it may lead to sidestep encounters with friends and family, in the end leading humans to no longer privilege authentic emotion, as warned by Turkle (2007). In the case of dependent people there is a symmetrical risk, namely that of allowing friends and family to sidestep their responsibilities. Turkle (2007) touches again on a far-reaching issue when she states, «the question is not whether children will love their robotic pets more than their animal pets, but rather, what loving will come to mean».

The decay of emotions is a recurrent theme throughout the novel. ***Silvana***, an `emotional masseuse´ that tries to help people recover lost sensations and reads old books to research the power of emotion, sees ***Celia*** as a living example of the feelings that are extinct at the time. Particularly in the highlights from Chapter 25, ***Silvana*** criticizes that ergonomically-designed technology discourages social relationship by preventing people from talking to one another along trips, and she strongly argues against robots being built that spoil, corrupt and brainwash people, hogging their most intimate time and space, so that they end up becoming wooden like them.

Question 5.C – Could robots be used as therapists for the mentally disabled?

Some psychologists suggest that the illusion of emotional understanding by a robot that makes eye contact and responds to touch may be therapeutic in some contexts. Additional virtues of robots as therapists are their endless "patience," their capacity for repetitive action without getting "bored," and their never showing unintended feelings, which some humans cannot repress.

Actually, interacting with robots that display social behavior has been shown to help children with autism acquire social skills [Feil-Seifer and Mataric 2008; Robins *et al.* 2005]. Although the goal of therapy is not to develop an attachment to the robot, it may occur as a side effect and, therefore, the ethical correctness of encouraging such children to engage in affective interactions with machines incapable of emotions is debatable. Whether the finding that severely autistic children prefer featureless, non human-like robots during play [Robins *et al.* 2004] should be interpreted in favor or against is unclear.

Further to the illusion of emotions discussed above, Turkle (2007) states, «If a person feels understood by an object lacking sentience, that makes eye contact and responds to touch, can that illusion of understanding be therapeutic?» and she continues to ask, «What is the status—therapeutic, moral, and relational—of the simulation of understanding?»

It is worth mentioning that robot-assisted therapy has been applied to other types of patients, such as diabetic children [Lewis *et al.* 2015; Nalin *et al.* 2012], with different aims to those for autistic patients: among them, reducing child's stress and anxiety, improving their response to medical treatments, promoting their self-efficacy, and encouraging physical activity. The use of robots in this context raises fewer doubts.

Nonetheless, Riek and Howard (2014) ask, «what happens when the therapy ends and the robot goes away?» Due to possible affective bonds with the robot, its disappearance may have counterproductive effects on the patient, even reversing the benefits of treatment. Thus, these authors suggest that the benefits and risks must be evaluated in advance and protocols must be specified for addressing this circumstance.

Question 5.D – How adaptive/tunable should robots be? Are there limits to human enhancement by robots?

There are two related issues here: up to what extent users should be able to (i) tune robot (possibly, moral) behavior and (ii) enhance themselves by means of robotic prostheses. As regards to the former, it seems clear that, for example, parents should

be able to modify the off-the-shelf robot skills to comply with their family values, or caregivers should be able to adapt a robot assistant to the particular needs of a patient. But are there frontiers that such tuning cannot trespass? Surely there are, as robots must be prevented from inflicting (physical or psychical) harm to people interacting with them, but setting up the limits is not an easy task.

Turning to the second issue, robotic devices can restore human sensing and physical mobility, thus helping to rebuild body image and restore performance, but they can go beyond that, leading to "human enhancement", i.e., improving human functions beyond what is necessary to sustain and reestablish good health. Again, establishing the limits is tricky: a wearable exoskeleton connected to the spinal cord of a stroke patient may restore their walking ability, and artificial retinas may palliate visual deficiencies, but it is not hard to imagine other uses of bio-robotic prostheses that may turn a human into a cyborg or a living weapon, maybe even remotely controlled by someone else. This extends to cognitive enhancement as well. One of the main themes of the novel is ***Dr. Craft***'s determination to get (and keep only for himself) a "creativity prosthesis" that enhances his inventive capacity, and ***Leo*** is in charge of developing it.

The debate is ultimately polarized into two main positions: transhumanists and bioconservatives. Transhumanism holds that the current form of the human species, on both somatic and cognitive levels, is merely a specific stage of human evolution, and we have only begun to grasp the extent of possible future integrations between the natural and artificial. Bioconservatism stresses the need to investigate the significance and the implications of the transformations concealed behind the apparently neutral technological development involving humans, thus placing the concepts of nature and human dignity as insurmountable limits [Palmerini *et al.* 2016].

The challenge is how to ensure that robots improve the quality of our daily lives, widen our capabilities, and increase our freedom, while avoiding their making us more dependent and emotionally weak; that is, the eternal dilemma of how to take the good without suffering from the bad side-effects. In their heated discussions, ***Leo*** defends the positive view of robots as enhancers of our physical and cognitive capabilities,

while ***Silvana*** highlights the downside that relating to robots ends up replacing people's intimate relationships.

5.3. Revisiting issues

Concerns raised in this section related to robots simulating emotions, thus possibly encouraging their users' affective attachment, were previously discussed under Questions 2.B and 2.D; and the possibility that this could lead to deception was also dealt with in Section 1.

The second highlighted episode from Chapter 25, in which ***Leo*** admits he is a guinea pig, permits revisiting Question 3.C. Moreover, Chapter 28 provides more details on the workings and implications of the time-out device discussed also in Section 3.

References

Boden M., Bryson J., Caldwell D., Dautenhahn K., Edwards L., Kember S., Newman P., Parry V., Pegman G., Rodden T., Sorrell T., Wallis M., Whitby B. and Winfield A.F. (2017) Principles of robotics: Regulating robots in the real world. *Connection Science*, 29(2): 124-129.

Bryson J.J. (2010) Robots should be slaves. In Wilks Y. (ed.): *Close engagements with artificial companions: key social, psychological, ethical and design issues*, pp. 63-74, Amsterdam, The Netherlands: John Benjamins Publishing Company.

Cheok A.D., Levy D., Karunanayaka K., Morisawa Y. (2017) Love and Sex with Robots. *Handbook of Digital Games and Entertainment Technologies*, Springer Singapore, pp. 833-858.

Cowie R. (2015) Ethical issues in affective computing. The Oxford handbook of affective computing, 334.

Feil-Seifer D., Mataric M. (2008) Robot-assisted therapy for children with autism spectrum disorders. Proc. 7th Intl. Conference on Interaction Design and Children, pp. 49–52.

Levy D. (2007) *Love, Sex with Robots: The Evolution of Human-Robot Relationships*, Harper Collins Publishing: New York.

Lewis M., Oleari E., Pozzi C., Cañamero L. (2015) An Embodied AI Approach to Individual Differences: Supporting Self-Efficacy in Diabetic Children with an Autonomous Robot. Tapus A., André E., Martin J-C., Ferland F., Ammi M. (eds.): *Social Robotics*. Lecture Notes in Computer Science, 9388, pp. 401-410, Springer.

Lichocki P., Kahn Jr P.H., Billard A. (2011) A survey of the robotics ethical landscape. *IEEE Robotics and Automation Magazine*, 18(1), 39-50.

Nalin M., Baroni I., Sanna A., Pozzi C. (2012) Robotic companion for diabetic children: emotional and educational support to diabetic children, through an interactive robot. Proc.

11th ACM Intl. Conference on Interaction Design and Children (IDC'12), New York, pp. 260–263.

Palmerini E., Azzarri F., Battaglia F., Bertolini A., Carnevale A., Carpaneto J., Cavallo F., Carlo A.D., Cempini M., Controzzi M., Koops B.J., Lucivero F., Mukerji N., Nocco L., Pirni A., Shah H., Salvini P., Schellekens M., Warwick K. (2016) Robolaw: Guidelines on Regulating Robotics.

Riek L., Howard D. (2014) A Code of Ethics for the Human-Robot Interaction Profession. *Proc. We Robot*, pp. 1-10.

Robins B., Dautenhahn K., Te Boerkhorst R., Billard A. (2004) Robots as assistive technology - does appearance matter? *13th IEEE Intl. Workshop on Robot and Human Interactive Communication (RO-MAN)*, pp. 277-282.

Robins B., Dautenhahn K., Te Boerkhorst R., Billard A. (2005) Robotic assistants in therapy and education of children with autism: Can a small humanoid robot help encourage social interaction skills? *Universal Access in the Information Society*, 4(2), 105–120.

Sharkey A., Sharkey N. (2012) Granny and the robots: ethical issues in robot care for the elderly. *Ethics and information technology*, 14(1), 27-40.

Sullins J. (2012) Robots, love, and sex: the ethics of building a love machine. *IEEE Trans. on Affective Comput*ing, 3(4), 398–409.

Turkle S. (2007) Authenticity in the age of digital companions. *Interaction Studies*, 8(3): 501-517.

6. Social responsibility and robot morality

6.1. Highlights from *The Vestigial Heart*

Chapter 29, page 214:

[Leo]'s ended up with not only his hands but also his brain tied to a company, and worse still, to its shady president. [..] it's undeniable that we're contributing to a veritable mutation of the species. Or rather, causing it. He looks at his hands as though he expects to find them more powerful, and stained. They're an extension of the Doctor, many hands like these forged the multitude of robots that exist around the world today, sculpting human nature. So much hidden power behind apparently loyal and useful servants.

Chapter 29, page 216:

[Celia:] "Now, though, I can't move an inch if it's not in an aero'car, and, of course, with ROBbie." [..] "It's not that I'm complaining about having him, he's an excellent toy, but having him watching over me all the time is a real pain in the neck."

Chapter 30, pages 225, 228-230:

[Dr. Craft:] "I don't merely want to benefit from his creativity, I want to expand my own!" [..] "Hook me up, that's an order."

[..]

[Alpha+:] 5:03 p.m. – These accessories have not been approved by the standards agency. I have to maximize precautions in order to avoid a severe penalization. [..] Even though ROBco advised me to monitor only the Doctor's basic variables, I will keep track of all his vital signs. As soon as one deviates from its baseline I will halt everything. I should not take any risks. More important than the whims of my PROP, I must safeguard his health.

[..]

[ROBco:] "Why have you connected sensors to his chest and the back of his neck? I did not tell you to."

5:08 p.m. – "I must ensure that the Doctor is not in danger at any moment."

"Acceptance: *It is your PROP. But it is also necessary to avoid him feeling uncomfortable."*

"Well said! Finally a ROB that's learned what it had to learn. [..] Come on, get all this stuff off me and turn the booth on once and for all, I want to try it."

5:09 p.m. – "Stop right there! Do not touch anything while the responsibility is mine."

"How dare you contradict me, you foul creature? You'll take it all off yourself, and I don't want to hear another word on the subject!"

5:10 p.m. – "Agreed." It obediently starts to remove the sensors. "But we will not be performing the experiment."

"What do you think you are, you useless bastard? I'm the one who makes the decisions. I don't need you, understand? Not for anything. Get the hell out of here before I immobilize you for good."

5:11 p.m. – "I object: that would be against the rules. I cannot abandon my PROP when he is in danger."

"Danger?" He stands up like a man possessed and heads for the robot. "You're the one who's become a danger: you drug me, you ration my pleasures, and now you want to prevent me from expanding my mind? It's over, you lump of scrap!"

5:12 p.m. – "What are you doing? Do not switch off my synthesizer. We can talk about this. I will help you get what you want."

"Not just the fucking synthesizer, no! I'll disconnect you completely this time ... and then I'll be able to live in peace!"

5:13 p.m. – "Careful, Doctor, everything has been recorded ... you know that Mr. Gat"

"There, fuck it, it's done."

He sits down again, satisfied, and turns to ROBco:

"Now, you, connect me to the bare essentials required to have my mind expanded just like your PROP's was."

Chapter 30, pages 231-232:

[ROBco reports to Leo:] When he was connected to the booth, the Doctor's vital signs strayed a long way from their baseline and the emergency protocol had to be applied. His recovery is moving at such a slow pace that the robot fears he could enter cardiac arrest at any moment and wants to know what effects suddenly stopping the session would have.

Leo jumps up as if he's received an electric shock and shouts: "Don't do it! It might kill him!" and starts pacing around the cubicle like an electron in a particle accelerator. He should have foreseen this, he thinks, the Doctor is an old man and his organs, which are accustomed to today's lifestyle, have lost their capacity to absorb strong emotions.

[..] ROBco insists:

"Warning: forty beats per minute, danger of cardio-respiratory arrest."

"What are you talking about? What's his ROB doing? It should be doing something!"

"Information: He disconnected it."

"WHAT??"

Leo drops into his seat dejectedly, and Celia takes his hand, as if she were comforting a sick person.

"Announcement: *The Doctor is dead. Question: What should I do?"*

"A death trap ... that's what I've invented. Now I'll have to go into hiding. What must you think of me, Silvana? You almost tried it out yourself ..."

She is momentarily paralyzed by the thought of what might have happened to her, but hearing the boy speak makes something inside of her rise up:

"Don't talk like that, it was an accident, it's not your fault. He was the one who disconnected his ROB, right? Maybe he knew exactly what he was getting himself into and that's what he wanted: to commit suicide."

"Much the opposite, he wanted to get younger, to suck the life out of someone else"—his eyes wander toward Celia, but he avoids looking at her. "Shame on me, I've been happily toying with the most delicate material in the world."

"Repetition: What should I do?"

"You two can tell it. I don't even know what to do with myself."

"Let's take this step-by-step." Silvana switches into crisis-management mode. "There must be someone we have to inform about what's happened."

"Yes, Mr. Gatew ... but they'll blame me ..."

"Clarification: The lady is right. They can't blame you because Alpha+'s record will have saved proof that the PROP disconnected it."

Chapter 30, page 234:

[Silvana:] "... whether I like it or not, robots have become the pro-technos' teachers, and we're better off letting them help people grow and become more creative than making people dependent and unimaginative.

[Leo:] "... There's nothing I'd like more than to make the creativity stimulator available to everyone, to put it on the public register." He smiles at Celia, he's willing to take that risk for her.

Chapter 31, page 235:

[Leo] was working on a top-secret device for his boss, I understood that much, and now that he's dead, he wants everyone to have one. Sounds easy doesn't it? Well it's not. To start with, he has to spend a long time locked up in the lab, like being kidnapped, and working twice as hard: for the new boss and, secretly, for all the people

he'll give the device to. He promised me that ROBbie would get the first one. Because, I haven't told you yet, the prosthesis—that's what they call it—has to be installed in a robot and is used to increase its owner's intelligence, if they want it to. The truth is, I doubt Lu or Fi would be interested in it, but Leo insists on putting the invention on the public register ...

6.2. Ethical Background and Discussion

Autonomous robots need to make decisions in situations unforeseen by their designers, which raises not only issues of reliability and safety for users, but also the challenge of regulating automatic decision-making, particularly in ethics-sensitive contexts. This has led to the development of the field of machine ethics, with the ultimate goal of coming up with methodologies for maximizing the likelihood that a robot will behave in a certifiably ethical fashion [Lichocki *et al.* 2011].

Some argue that robots can be better moral decision makers than humans, since their rationality is not limited by jealousy, fear, or emotional blackmail [Wallach 2010], whereas others argue that machines can never be moral agents and, therefore, they should not be endowed with the capability of making moral decisions.

Even assuming that general ethics rules could be implemented in robots, questions then arise as to who should decide what morality is to be encoded in such rules and up to what point the rules should be modifiable by the user. For instance, it is unclear whether and when it may be acceptable to intrude upon a robot user's autonomy to become "more ethical" towards other human beings or in the hope of making other people's life better [Borenstein and Arkin 2016].

In any case, a robot is a tool and, as such, it is never legally responsible for anything. Therefore, it is of utmost importance to establish procedures for attributing responsibility for robots, so that it will always be possible to determine who is legally responsible for their actions [Boden *et al.* 2017]. In the case of robots able to learn from experience, such responsibility may be shared between the designer, the manufacturer, and the user; a hacker may also be charged with it if their illegal intervention can be demonstrated.

In Chapter 30, ***Alpha+*** says it is against the rules to abandon its PROP while he is in danger. But its PROP, ***Dr. Craft***, is ultimately the one who decides and switches his robot off. Who is responsible for the fatal consequences? ***Leo*** feels doubly guilty, as designer of the sensory booth—a "death trap," he calls it—and as the PROP of ***ROBco***, the robot directly involved in the death, whereas ***Silvana*** claims that it was either an accident or a suicide.

Question 6.A – Can reliability/safety be guaranteed? How can hacking/vandalism be prevented?

No computational system can be proven to be entirely error-free or vandal-proof under all circumstances. However, more and more sophisticated robot safety and security measures are being developed and standards are being established and enforced by competent agencies, such as the Robotic Industries Association [RIA] and the IEEE Standards Association [IEEE SA 2017].

In the first highlights from Chapter 30, ***Dr. Craft*** asks ***Alpha+*** to connect him to accessories that have not been approved by the standards agency, thus in order to safeguard the doctor's health, the robot sets up to maximize precautions. This illustrates the robot adhering to the *precautionary principle*: «When an activity raises threats of harm to human health or the environment, precautionary measures should be taken even if some cause-and-effect relationships are not fully established scientifically», which all professionals are advised to apply in dealing with sensitive technologies [Veruggio *et al.* 2016]. ***Alpha+*** tries not only to perform safe actions, but also to ensure the safety of its PROP under the action of others by refusing to leave its PROP when he may be in danger.

Even if robots are designed to be safe and secure, users or hackers may make them do things their designers did not foresee. Recall that, in Chapter 5, ***Leo*** modifies ***ROBco***'s software in a way that contravenes manufacturing specifications and safety rules. Regulations must establish how far those who own or operate robots should be required or allowed to protect them from e.g. bad praxis, theft, or vandalism [Boden *et al.* 2017].

Leroux and Labruto (2012) consider the question of whether a "human-in-the-control-loop" requirement should be enforced without exception. This may affect safety in positive and negative ways. For example, in shared-control systems, provisions need to be made to prevent human habituation to automatic functioning, so that the person doesn't become bored or distracted, thus disregarding their duties. This could be implemented through preplanned episodes of handoff to the human controller for the purpose of maintaining human attention and skill levels.

Question 6.B – Who is responsible for the actions of robots? Should moral behavior be modifiable in robots?

«A world without consequences and costs is a world without meaningful choice. A life without responsibility is not the life of the adult –it's the life of the animal, the child, or the robot.» [Roberts 2001]. Most roboticists would agree with this quotation from a fiction book that attributes responsibility exclusively to adult humans. However, such attribution becomes increasingly complex as robots become more autonomous and capable of modifying their behavior through learning and experience, since their actuation is no longer based entirely on their original design.

Until now, if a machine went wrong it was always the maker or the programmer—or their company—which was at fault. With the advent of learning robots, a grey area of responsibility embraces those above together with the owner and the user; and a proposal was put forth by Decker (2007) «that robot learning should be anchored in the liability of the robot's owner», as derived from Kant's formula of humanity. Note that this author refers to liability, which is the legal consequence of responsibility. Along this line and in accordance with the quotation above, it has been suggested that the liability of animal keepers could be used as a model for the liability of robot keepers [Schaerer *et al.* 2009].

Another option is meta-regulation by a liability arbitration institution. For litigation purposes, a robot's decision path would then need to be traceable, a possibility being

to install a non-manipulable "black box" to continuously document the significant results of the learning process and the relevant inputs, which could be checked by the aforementioned institution. To convince ***Leo*** that he cannot be blamed for ***Dr. Craft***'s death, ***ROBco*** reminds him that ***Alpha+***'s record will have saved proof that its PROP disconnected it.

Manufacturers could protect themselves from liability by asking the robot owner to confirm, for example by pressing a button, that the robot learning process has been made transparent and he agrees to it. This confirmation would be recorded in the black box, and liability would be placed on the owner's side, as proposed by Decker (2007). The robot manufacturer would only need to refer clearly in the instructions to this confirmation procedure and to it being recorded in the black box.

Peltu and Wilks (2010) envisage even another possibility, namely that technology developments influence changes in the law, so that things that are not human, such as robots, could be liable for damages.

Question 6.C – When should a society's well-being prevail over the privacy of personal data?

This question often arises in the medical context, where the social importance of collecting data for research purposes may get into conflict with the patients' right to privacy. Following the precautionary principle mentioned under Question 6.A, data protection procedures have been developed and the use of informed consent forms has been encouraged. As people increasingly interact with robots in a social context (e.g, in the role of sales agents, health carers, or similar assistants), the risk of unintended (or intended) information disclosure and its use for commercial purposes increases.

Calo (2015) describes the ways in which cyberlaw developed for the Internet need to be extended to cover additional issues raised by social robots. For example, a robot introduced into the home could compromise privacy merely by creating the sense of being observed. This concern appears in Chapter 29, when ***Celia*** complains about her

being watched all the time by ***ROBbie***. But the uncomfortable sensation may turn into real danger if vacuum cleaners, window washers, child companions, and assistants to the elderly and disabled could become spies, especially if hacked by third parties.

The other highlighted episode from Chapter 29, in which ***Leo*** feels guilty of having forged robots that are sculpting human nature in undesirable ways, raises a more abstract, far-reaching concern about human evolution and society's well-being in a progressively robotized world. This appeal to social responsibility underlies the inquiry discussed by Borenstein and Arkin (2016): «Does the foremost obligation that a robot possesses belong to its owner or to human society overall?» As these authors warn, the answer to this question can have a profound impact on the robot's design architecture.

Question 6.D – What digital divides may robotics cause?

It is well known that digital technologies open up important social divides (based on age, wealth, education, world areas) and robots may widen some of these because of their cost, physical embodiment, and nontrivial usage [Veruggio *et al.* 2016].

An example of divide per age, education, or simply individual preference, is when a citizen can only get a service by interacting with a robotic agent. Regulations must guarantee the right of everybody to equalitarian access to services and, thus, the option of being redirected to a human agent should always be in place.

Technology has a strong impact on the global distribution of wealth and power, causing divides per world areas. Nagenborg *et al.* (2008) make the point that «the effects of the increasing use of robots in the world of work cannot be judged only by looking at those countries where these robots are used. There must also be questioning about the effects on other countries (brain drain, loss of jobs, etc.) and the relationship between countries that might be affected by what they call the *robotic divide*».

Conversely, robotic assistants targeted at vulnerable groups could reduce social discriminations and help shrink the aforementioned divides if policy measures were

taken to provide the required financial resources and knowhow to such groups [Peltu and Wilks 2010]. The last highlighted episode from Chapter 30 shows that ***Leo*** is aware of this social problem and decides to sacrifice his immediate freedom to work toward making the creativity prosthesis available to everybody. In Chapter 31, ***Celia*** tells her mother how proud she is that he is willing to do so, even if people like ***Lu*** and ***Fi*** may not be interested in the benefits such a prosthesis could bring them.

6.3. Revisiting issues

In Chapter 29 ***Leo*** refers to the timeout device as a way of renting brains, an enslavement mechanism that violates the rights of employees, thus permitting to deepen in some of the issues discussed under Question 3.D.

Moreover, in Chapter 30, ***Leo*** worries that ***Dr. Craft***'s organs may have lost the capacity to absorb strong emotions, implying that emotions have disappeared due to the lifestyle prevailing in their robotized society. This may lead to revisiting the trade-off discussed in Section 5 that our close interaction with robots may widen some of our capabilities, but at the risk of making us emotionally weak.

References

Boden M., Bryson J., Caldwell D., Dautenhahn K., Edwards L., Kember S., Newman P., Parry V., Pegman G., Rodden T., Sorrell T., Wallis M., Whitby B. and Winfield A.F. (2017) Principles of robotics: Regulating robots in the real world. *Connection Science*, 29(2): 124-129.

Borenstein J., Arkin R. (2016) Robotic nudges: the ethics of engineering a more socially just human being. *Science and Engineering Ethics*, 22(1), 31-46.

Calo R. (2015) Robotics and the Lessons of Cyberlaw. *California Law Review*, 103(3), 513-563.

Decker M. (2007) Can humans be replaced by autonomous robots? Ethical reflections in the framework of an interdisciplinary technology assessment, *Workshop on Roboethics*, Intl. Conf. on Robotics and Automation (ICRA'07).

IEEE SA (2017) Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems. Online: https://standards.ieee.org/develop/indconn/ec/autonomous_systems.html

Leroux C., Labruto R. (2012) Ethical, Legal, and Societal Issues in Robotics, euRobotics: The European Robotics Coordination Action, Deliverable D3.2.1.

Lichocki P., Kahn Jr P.H., Billard A. (2011) A survey of the robotics ethical landscape. *IEEE Robotics and Automation Magazine*, 18(1), 39-50.

Nagenborg M., Capurro R., Weber J., Pingel C. (2008) Ethical regulations on robotics in Europe. *AI & Society*, 22(3), 349-366.

Peltu M. and Wilks Y. (2010) In Wilks Y. (ed.): Summary and discussion of the issues. In Wilks Y. (Ed.) *Close engagements with artificial companions: key social, psychological, ethical and design issues*, pp. 259-286, Amsterdam, The Netherlands: John Benjamins Publishing Company.

RIA https://www.robotics.org/robotic-standards

Roberts R. (2001) *The Invisible Heart - An Economic Romance*, MIT Press.

Schaerer E., Kelley R., Nicolescu M. (2009) Robots as animals: A framework for liability and responsibility in human-robot interactions. Proc. 18th IEEE Intl. Symposium on Robot and Human Interactive Communication (RO-MAN), pp. 72-77.

Veruggio G., Operto F., Bekey G. (2016) Roboethics: Social and ethical implications of robotics. In Siciliano B., Khatib O. (Eds.) *Springer Handbook of Robotics*, *2nd edition*, Chapter 80, pp. 2135-2160, Springer.

Wallach W. (2010) Robot Morals and Human Ethics: The Seminar, *Teaching Ethics*, 11(1), 87-92.

7. Bibliography and Initiatives to Follow up

7.1. Roboethics Books Complementing our Focus on Social Robotics

Since Roboethics is a relatively new subject, there are few books covering its scope, and even fewer intended as textbooks. Perhaps the most representative is that edited by Lin *et al.* (2011), a nice collection of 22 essays by 27 contributors, covering most of the relevant topics, such as military issues, law, medicine, sex, emotional bonds, privacy, liability and moral agency. This wide span and the very authoritative authors make of this book a reference treatise, especially in the humanities, given its mostly philosophical stance.

In this teacher's guide, we have focused on social robots and the practical concerns they raise for engineers and users. Since our proposal is thus limited in both scope and orientation, we next point to some references where the left out aspects are thoroughly addressed.

Concerning scope, some books centered on a roboethics topic complementary to ours are those by Arkin (2009), Singer (2009) and Krishnan (2009) on the military, Pagallo (2013) and Calo *et al.* (2016) on law, van Rysewyk and Pontier (2015) and van Wynsberghe (2015) on medical robots, and Levy (2009) on love and sex with robots.

Books with a predominant philosophical orientation are those by Wallach and Allen (2008), Anderson and Anderson (2011), Gunkel (2012), Dekker and Gutmann (2012), and the forthcoming one by Wallach and Asaro (2016), whose primary aim is to develop a code of ethics for machines, thus dealing with speculative issues such as endowing robots with consciousness and morality.

A book that, like this guide, recurs to science-fiction stories for illustration is that by Nourbakhsh (2013), which by drawing some possible future scenarios raises some concerns about where we are heading, without neither taking an ethics viewpoint nor explicitly trying to be pedagogical. The author, a renowned roboticist, makes very lucid

remarks by concentrating on just six specific topics (marketing strategies in the net, the consequences of non-ephemeral design, etc.), some barely related to robotics, though.

Two books that have a technological orientation close to ours are those by Capurro and Nagenborg (2009) and Tzafestas (2015). The former is an edited volume focusing on the practical ethical questions raised by human-robot interaction, from the perspective of humans and taking into account intercultural differences. The latter is a single-authored book starting with a very helpful bibliography overview and touching on every topic related to roboethics with a commendable encyclopedic spirit. Both books devote major attention to the currently most eye-catching ethical issues, namely warfare, medical and social robotics.

7.2. Robotics Meets the Humanities

The study of the ethical implications of social robotics calls for the collaboration between researchers in robotics and the humanities. Many joint initiatives have emerged, such as the organization of regular workshops and seminars (e.g., [ICRA Forum 2013]), the publication of special issues in scientific journals [Veruggio *et al.* 2011], the launching of research projects [euRobotics 2012; RoboLaw 2014] and open discussion forums such as the Open Roboethics Initiative. This is a web space where policy makers, engineers, designers, users and other stakeholders of the technology can freely share and access roboethics related contents, with the aim to accelerate discussions and inform robot designs. Moon *et al.* (2012) motivate the need for such an initiative and provide an overview of the short history of Roboethics.

Furthermore, professional associations have launched initiatives for developing ethical codes in robotics and intelligent systems; examples are those put forth by the IEEE Standards Association and the Bristish Standards Institution. Likewise, the European Parliament has released some guidelines under the general title of “Civil Law Rules on Robotics”.

7.3. Roboethics Initiatives based on Science Fiction

The current acceleration of technological development makes it difficult to scientifically predict how our increasing interaction with robots will affect the evolution of society, the economy, and the life of people in a few years time [Torras 2016]. However, imagining possible future scenarios is what science fiction is best at, as offered by Stephenson (2011) in his thought-provoking talk. Actually, Asimov (1978) anticipated today's state of affairs when he stated that change is the dominant factor in society and «our statesmen, our businessmen, our everymen must take on *a science fictional way of thinking*».

With this aim, several initiatives have recurred to science fiction to explore possible benefits and risks of some technological innovations [Torras 2015]. An example is *The Tomorrow Project*, an initiative launched by the company Intel, in which four science-fiction authors were asked to write short stories about the future use of their products in photonics, robotics, telematics and smart sensors [Rushkoff *et al.* 2012] and that has been continued at the Center for Science and the Imagination of Arizona State University.

Thus, when trying to establish an ethical debate, disseminate concepts to the general public, or teach a course on roboethics, science fiction is often used to exemplify possible future conflicting situations, as we did here with *The Vestigial Heart*. Along this line, themes addressed in the classic works by Asimov, Dick, Bradbury, Capek, Shelley, or Hoffman, such as the three laws of robotics, robot nannies, humanoid replicas, or emotional surrogates have attained great relevance with the development of social robots.

Asimov's three laws of robotics, simplistic and generally impractical as they may be, have proven provocative and useful to trigger ethical research [Anderson 2008; Murphy and Woods 2009]. Most concerns about robot nannies discussed in Section 4 already appeared in science fiction stories published more than half a century ago, [Asimov 1950; Dick 1955; Bradbury 1969], such as the protection/freedom dilemma, the lack of privacy, the risk of deceit, and the difficulty of acquiring capacities like empathy and autonomy [Torras 2010]. Remarkably much earlier, Hoffmann (1816)

illustrated the psychological and social problems that emotional attachment to an artificial creature could cause, and Shelley (1818) dealt also with the emotions and consequences of giving birth to a human-like being by scientific means. Likewise, Capek (1920) anticipated many of the issues raised by humanoid robots in relation to anthropomorphism, impact on employment, human-robot interaction and social responsibility [Christoforou and Müller 2016], which have been discussed in Sections 2, 3, 5 and 6, respectively.

Modern science fiction literature touches on many of the ethical concerns here considered. Two novels that complement our treatment in dealing with far longer-term issues are those by Bacigalupi (2009), about a robot developing consciousness of having been built to serve, and by Chiang (2010), showing the problems that attachment to artificial pets could produce.

Given the rising interest in these speculative themes nowadays, many 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 the films like *Blade Runner 2049*, *Surrogates* and *Robot and Frank*. Actually, the latter is being used in the platform Teach with Movies 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.

To conclude, it seems clear that social robots point to some nuanced social issues and pose intriguing ethical questions, which open up amazing possibilities for the future. In this very delicate area, science fiction may help us clarify the role that the human

being and the robot have to play in this *pas de deux* in which we are irrevocably engaged.

References

Anderson S.L. (2008) Asimov's "three laws of robotics" and machine metaethics. *AI & Society*, 22(4), 477-493.

Anderson M., Anderson S.L. (Eds.) (2011) *Machine ethics*. Cambridge University Press.

Arkin R. (2009) *Governing lethal behavior in autonomous robots*. CRC Press.

Asimov I. (1950) Robbie. In: *I, Robot*. New York: Gnome Press.

Asimov I. (1978) My own view. R. Holdstock (Ed.) *The Encyclopedia of Science Fiction*, 5. New York: Saint Martin's Press.

Bacigalupi P. (2009) *The Windup Girl*, Night Shade Books.

Bradbury R. (1969) *I sing the body electric*. New York: Knopf Publishing Co.

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

Capek K. (1920) *R.U.R. Rossum's Universal Robots*.

Calo R., Froomkin M., Kerr I. (Eds.) (2016) *Robot Law*. Edward Elgar: Northampton.

Capurro R., Nagenborg M. (Eds.) (2009) *Ethics and robotics*. IOS Press.

Center for Science and the Imagination, Arizona State University. http://csi.asu.edu/category/project-archive/tomorrow-project/

Chiang T. (2010) *The Lifecycle of Software Objects*, Subterranean Press.

Christoforou E.G., Müller A. (2016). RUR revisited: perspectives and reflections on modern robotics. *Intl. Journal of Social Robotics*, 8(2), 237-246.

Dekker M., Gutmann M. (2012) Robo- and informationethics. *Some fundamentals. LIT-Verlag, Wien*.

Dick Ph.K. (1955) Nanny. *Startling Stories*, Spring issue.

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

Gunkel D.J. (2012) *The machine question: critical perspectives on AI, robots, and ethics*. MIT Press.

Hoffmann E.T.A. (1816) *The Sandman*.

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

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

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.

Krishnan A. (2009) *Killer robots: legality and ethicality of autonomous weapons*. Ashgate Publishing.

Levy D. (2009) *Love and Sex with Robots: The Evolution of Human-Robot Relationships*, Harper Collins Publishing: New York.

Lin P., Abney K., Bekey G.A. (2011) *Robot ethics: the ethical and social implications of robotics*. MIT Press.

Moon A., Calisgan E., Operto F., Veruggio G., Van der Loos H.M. (2012) Open Roboethics: Establishing an Online Community for Accelerated Policy and Design Change. *We Robot Conference*.

Murphy R., Woods D.D. (2009) Beyond Asimov: The three laws of responsible robotics. *IEEE Intelligent Systems*, 24(4), 14-20.

Nourbakhsh I.R. (2013) *Robot futures*. MIT Press.

Open Roboethics Initiative http://www.openroboethics.org/

Pagallo U. (2013) *The laws of robots: crimes, contracts, and torts* (Vol. 10). Law, Governance and Technology Series 1, Springer.

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.

Shelley M. (1818) *Frankenstein or The Modern Prometeus*.

Singer P.W. (2009) *Wired for War: The Robotics Revolution and Conflict in the 21st Century*, Penguin Press.

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

Teach with Movies. Robot Ethics Using Clips from *Robot and Frank.* http://www.teachwithmovies.org/snippets/sn-sci-robot-ethics-robot-and-frank.html

Torras C. (2010) Robbie, the pioneer robot nanny: Science fiction helps develop ethical social opinion. *Interaction Studies*, 11(2), 269-273.

Torras C. (2015) Social robots: A meeting point between science and fiction. *Metode Science Studies Journal-Annual Review*, 5, 111-115. Available online: http://www.redalyc.org/pdf/5117/511751360016.pdf

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

Tzafestas S.G. (2015). *Roboethics: A Navigating Overview*. Intelligent Systems, Control and Automation: Science and Engineering Series 1046. Springer.

van Rysewyk S.P., Pontier M. (2015) *Machine Medical Ethics*. Springer.

van Wynsberghe A. (2015) *Healthcare Robots: Ethics, Design and Implementation*. Ashgate Publishing.

Veruggio G., Solis J., Van der Loos M. (Eds.) (2011) Roboethics: Defining Responsibility to Protect Humankind, Special Issue of the *IEEE Robotics and Automation Magazine*, 18(1).

Wallach W., Allen C. (2009) *Moral Machines: Teaching Robots Right From Wrong.* New York: Oxford University Press.

Wallach W., Asaro P. (Eds.) (2016) *Machine Ethics and Robot Ethics*. Ashgate Publishing.