Ethics of Social Robotics: Individual and Societal Concerns and Opportunities

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

ABSTRACT

Focus on the ethics of a given technology tends to lag far behind its development. This lag has been particularly acute in the case of Artificial Intelligence, whose accelerated deployment in wide activity sectors has triggered unprecedented attention on the risks and consequences for society at large, leading to a myriad of ethics regulations, which are difficult to coordinate and integrate due to their late appearance. The very nature of social robots forces their deployment to occur at a much slower pace, providing an opportunity for a profound reflection on ethics, which is already happening in multidisciplinary teams. This paper provides a personal view of the ethics landscape, centered on the particularities of social robotics, with the main issues being ordered along two axes (individual and societal) and grouped into eight categories (human dignity, human autonomy, robot transparency, emotional bonding, privacy and safety, justice, freedom, and responsibility). This structure stems from the experience in developing and teaching a university course on 'Ethics in Social Robotics´, whose pedagogical materials are freely available.

Keywords: robot ethics, social robotics, human-robot interaction, AI ethics, philosophy of technology, social responsibility

1. INTRODUCTION

Social robotics is a trending research topic nowadays as the embodiment of artificial intelligence in devices that can move and act in the real world is rapidly gaining presence in our daily lives (1). In view of the enormous disruptive potential of this technology, a profound ethical reflection on the deployment of social robots in human-centered environments has become not only necessary, but urgent. More so because, as usual, the technology is developing faster than its agreed-upon regulation, and several organizations and associations have felt the pressing need to develop standards and guidelines specific for their particular domains of activity.

The term ‘social robot’ has been used to refer to widely different entities in literature, ranging from on-screen avatars to all sorts of domain-specific service robots acting in human environments, potentially including even those deployed for warfare. This review will adhere to a more restrictive notion of social robot, namely an autonomous, embodied, artificially-intelligent agent that interacts and intentionally communicates with humans to provide a service in a social context. Prominent such contexts include healthcare, education, sales and entertainment.

The aim of this article is to provide an ethics overview very focused on social robots as just defined, but widely open in the temporal dimension. Ethics is a human endeavor with a long history and, the next section situates current developments in the framework of the philosophy of technology and the assumption of responsibility with respect to the evolution of humankind.

When having a look at the current state of affairs, we see that most works concentrate on the benefits and risks of human-robot interaction, i.e., they circumscribe the ethical analysis to the dyadic interaction of the robot with either the user or eventually a caregiver, lacking a more global systemic and societal perspective (2). This stems from the fact that the majority of these works are authored by robotics researchers, the same ones carrying out the technical work and designing the experiments. Recently, the situation has changed a bit with the incorporation of philosophers and social scientists into multidisciplinary research teams, and more global outlooks have begun to emerge. The main dyadic and societal issues that have appeared in the specialized literature on social robotics are reviewed In Sections 3 and 4, respectively.

It is worth mentioning that while initially most works showed strong concerns about the risks and possible negative consequences of interacting with social robots, today there is rising support for the optimistic claim that such interaction may provide opportunities for personal growth, new forms of relationship, and the development of individual and social values. Thus, the last section delves into these opportunities and draws some conclusions and lines along which to conduct future ethics research.

2. ETHICS HISTORICAL CONTEXT

Artificial is, by definition, anything made or produced by humans, especially as a copy of something natural. Social robots clearly fit this definition and, as the Nobel laureate Herbert A. Simon (3) nicely worded, they are the fruit of human imagination and the great technological development it subsequently brought about. Fictitious robots were forged as our mirrors; they reflected our dreams and our fears (4). Like a boomerang, they gave us back what we projected into them. Now we build them and endow them with cognition to interact with us. We've reached an inflection point in our joint history: social robots learn from us and in turn influence us. The loop closes and a new era opens. We are still at the very beginning, but technological acceleration and the keen adoption of automated assistance by social institutions will make everything evolve fast. Thus, we better turn to the long-standing philosophical ethics tradition and rely on well-founded ethics principles to reflect on the type of robots by which us and future generations are to be modeled.

Despite the antiquity of artificial products and the use of tools to build them, it is not until the beginning of the 20th century that philosophers such as Martin Heidegger (5) start to reflect on what they call 'the age of technique'. With this expression they want to make it clear that technology is transforming the natural world and human life in a much more radical and global way than it had done until then. In addition to trying to understand this historical specificity, they are driven by the urgency to analyze the ethical implications of technological deployment. In this context, Hans Jonas (6) establishes the well-known principle of responsibility of present humans towards future humans who should be able to consider this same question of responsibility.

With the spread of digital technologies, this process of transformation of human life has accelerated to the point where a qualitative leap has occurred, leading to what we could call 'the age of digitalization'. Artificiality has reached a new dimension and no longer affects only materiality and our physical abilities to act on reality, but extends to the sphere of psychic capacities that have been exclusively human, such as reasoning, adhering to values, artistic expression and decision-making. Industrial robots, which took on some physical tasks in manufacturing chains and production processes, have given way to social robots performing perceptive, cognitive and dexterous tasks we used to accomplish in our daily life.

If technique had already had an impact on natural ecosystems and Darwinian evolution, that is to say, in the domains of biology and medicine, now digital technologies are also disruptive in the field of thought, social habits and, ultimately, in the development of culture. And this disruption is occurring at the high speed characteristic of digital computing, far from the time scale in which people move. As Bilbeny (7) cleverly adverts: «What a human has learned —language, knowledge, emotional and social behavior— cannot be transmitted completely and faithfully to another human being. Culture is a constant redo. But a robot can transmit to another exactly everything it knows. Once again we see and suffer from the lag or lack of synchrony between the speed with which technology advances and the slowness and contradictions with which culture does so.»

Some anthropologists claim that we are at a crossroads in the evolution of humanity (8) and, as an intelligent and conscious species, we have the responsibility to avoid our extinction by stopping globalization, promoting diversity, and collectively opting for a new humanism based on the socialization of technology.

Focusing on the portion of responsibility attributable to robot deployment, let's start by defining robot ethics as the subarea 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. Two branches are often distinguished: human ethics applied to robotics, and codes of ethics embedded in the robots themselves (sometimes named "machine ethics") (9). We will here concentrate on the former and touch on the latter when relevant to the discussion of a particular issue.

Sullins (10) briefly surveys the main ethical theories relevant to robotics, namely:

- 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.
- Information ethics: policies and codes for governing the creation, organization, dissemination, and use of information.

Since no single theory is appropriate for addressing all ethical issues arising in the design and use of robots, Wallach & Allen (11) proposed a hybrid approach combining 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). This pragmatic approach is the prevalent one in the robotics community so far.

In order to structure the review of the ethical issues arising in social robotics, it is important to group them under fundamental categories. Institutions and researchers having undertaken the same task in the more general areas of robotics (12-14) and artificial intelligence (15-18) proposed categorizations with some commonalities, but lacking a clear consensus. Taking these into account and also the specificities of social robotics partly addressed by other authors (2, 19-22), Sections 3 and 4 review human-robot interaction issues and more global societal aspects, respectively, grouped in broad categories, which are not clear-cut but interrelated as will be noted in their descriptions below.

3. ETHICS OF HUMAN-ROBOT INTERACTION

Having human rights in mind and previous systematizations as just mentioned, we group dyadic ethics issues in five categories: human dignity, human autonomy, robot transparency, emotional bonding, and privacy and safety.

3.1. Human Dignity

There is consensus that social robots should be designed in ways that do not denigrate humans. This entails not only using respectful language and never intimidating users, but also having 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 (23). For example, in order to avoid eliciting feelings of objectification and loss of control, robots should not touch people or impose them some behaviors without previously informing and consulting them, i.e., robots should never reduce users to a machine-like status (24).

A feeling of vulnerability similar to that caused by an unforeseen physical contact with a robot may occur at the cognitive level, due to a mismatch in communication occasioned by the robot misinterpreting the user situation. Even in the restricted domain of automatic emotion detection, 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 (25) 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.

Furthermore, the useful capacity of robots to monitor and make decisions about a person’s health must be balanced with that person’s right to control over their own life (e.g., refusing treatment) (13). This raises questions as to the extent to which the wishes of a patient or elderly person must be followed, and the relationship between the amount of control given to them and their state of mind (26), which is difficult to evaluate and evolves over time.

In sum, 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. Ultimately, 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 possibility to do so, a guideline that is not easy to implement, as the many companies using chatbots to provide customer support demonstrate.

A related issue is 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. This will be discussed in the next section.

3.2. Human Autonomy

The key education dilemma between protecting and promoting autonomy in children appears also in the context of human-robot interaction. Besides the need for protection and help, addiction and manipulation can also compromise autonomy.

Sharkey & Sharkey (27) 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?» and Wilks (28) raises the question of how would children feel if their parents knew what they were doing all the time. These are extreme cases 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. Pearson & Borenstein (29) 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.

Turning to adults and specifically elderly groups, Espingardeiro (30) analyzes the thin boundary between comforting exercises and addiction to robots. Although design and manufacturing standards try to minimize the user's risk of addictions, Oliver (31) 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 (32).

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 good habits in their users, such as recycling for sustainability (33) or healthy practices 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. The latter is a societal aspect to be discussed in Section 4.2.

Borenstein & Arkin (34) refer to this robot tactic of subtly influencing its user as "nudging". Thaler & Sunstein (35) envisage three design pathways: "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?

The nudging potential of robots is higher than that of cellular phones and intelligent watches, because their autonomous motion permits following and monitoring the user, and their compelling physical presence is much more persuasive (36).

3.3. Robot Transparency

Winfield *et al.* (37) argue that transparency of autonomous and intelligent systems (AIS) is necessary: i) to discover how and why something went wrong, ii) to make accountability possible, and iii) for AIS to be understandable by users. While the first two items are related to safety and will be addressed in Section 3.5, here we focus on the last item, which has a specific significance for social robots, since their appearance and behavior may mislead users.

Although an agreed principle is that «robots should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be transparent» (13), the risk of deception in their deployment is high and takes many forms depending on context (38). 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. Whether deception is viewed as ethically acceptable depends very much on context and needs to be investigated in practice (39). For example, it may be perceived as correct if it increases benefits for the deceived and there is no betrayal of trust. Of course, some cases are more morally reprehensible than others (40), but in any case robots should not impersonate human agency by attempting to mimic intentional states.

The paradox, especially in the case of anthropomorphic robots, is that their design conveys human attributes, thus fostering this deceit problem. Moreover, as Breazeal adverts, «give hearing and voice to a robot and people expect it to be intelligent» (41). 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.

Riek and Howard (42) extend transparency to robot programming and predictability of future robot moves. Along this line, Van der Loos (43) suggests software developers should 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 through, for example, gestures, voice and context. This emphasis on communication is supported by a field study carried out by Dautenhahn *et al.* (44) on people's preferences as regards to assistant robots, where humanlike communication was largely prioritized over humanlike behavior and appearance.

3.4. Emotional Bonding

The idea of robot companionship seems natural to some people and almost obscene to others. Levy, in his provocative book (45) and in a review of the state of affairs ten years later (46), maintains that many people will no doubt fall in love with robots and that this is completely normal. On the other hand, Bryson (47) argues that artificial companions should just be servants, machines that you should be able to switch off whenever you like. Sullins (48) holds an intermediate position in proposing ethical design principles to limit the way emotions are simulated in robots.

Whether establishing emotional bonds with robots could improve the quality of life of some people or just create dependencies doesn't have a clear-cut answer, as it depends on the context and requires careful analysis (49). For example, Vallverdú & Casacuberta (50) argue that empathy is the key emotion in healthcare and that machines need to be able to detect and mimic it. They view the establishment of emotional bonds between humans and machines in a very positive way as the outcome of a global trust process in which emotions are co-created between machine and human.

However, the illusion of robot 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 (13). Turkle (51) alerts that, although the robot is only expressing a simulated emotion, the feelings it evokes in people are real and may be strong. A balance needs to be reached since, for instance, human caregivers sometimes simulate affection to improve their patient's wellbeing, and thus robots may also be allowed to do so under similar circumstances.

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 Turkle (51) states, «in the culture of simulation, authenticity is for us what sex was to the Victorians: taboo and fascination, threat and preoccupation.»

The risk that easy attachment to a robot would erode the person's motivation for engaging in human relationships, which may seem too hard, has been called the 'lotus eater' problem (25). 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. Turkle (51) 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».

In the case of dependent people there is also a symmetrical risk, namely that of allowing family and friends to sidestep their responsibilities once the care activities are covered by the robot, resulting in the user's social isolation.

Particularly in healthcare and educational contexts, it is clear that robots may relieve workers from some routine tasks with no special added value, but never replace them in their entire jobs. There are roles that can never be fulfilled by artificial agents, especially those entailing affection, life experience, and transmission of human values, i.e., those usually leading to emotional bonding.

A final note of caution: beyond the discussion of whether robot designs should encourage or discourage the formation of emotional bonds, roboticists must be aware that some bonding will be inevitable regardless of the morphology of the robot (52).

3.5. Privacy and Safety

Emotional bonding with a robot may raise also privacy concerns, as it may induce users to divulge more of their life and personal data than they would in a 'normal' setting (53). This relates also to transparency, since how social robots actually function and use data should be made clear to users, to avoid their being misled by a friendly appearance and simulated emotions.

Calo (54) describes the ways in which cyberlaw developed for the Internet needs 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. But the uncomfortable sensation may turn into real danger if vacuum cleaners, window washers, child companions, and assistants to the elderly could become spies, especially if hacked by third parties.

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 the *precautionary principle* should always be applied (55): «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».

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, together with the recognition of the mediating role of digital technologies (56), has led to design with embedded ethics (57), with the ultimate goal of coming up with methodologies for maximizing the likelihood that a robot will behave in a certifiably ethical fashion (24).

Some argue that robots can be better moral decision makers than humans, since their rationality is not limited by jealousy, fear, or emotional blackmail (58), whereas others argue that machines can never be moral agents and, therefore, they should not be endowed with the capability of making moral decisions.

Leroux & Labruto (59) 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.

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 liability for robots, so that accountability for their actions can be established based on the traceability and transparency of their behavior (13). For robots able to learn from experience, such liability may be shared between the designer, the manufacturer, the owner and whoever had interacted with it. Traceability is even more pressing in this case, and a suggested option has been to install a non-manipulable “black box” to continuously document the significant results of the learning process and the relevant inputs.

4. SOCIETAL ETHICAL ISSUES

As regards to more global concerns at the meso and macro levels, we consider three categories: justice, freedom, and responsibility, both at the professional level and as regards to future generations.

4.1. Justice

Some concerns have to do with the distribution of benefits and burdens across members of societies. For example, how to deploy social robots in a way that contributes to a fairer distribution of care (2). Tackling this concern at a meso level, van Wynsberghe & Li (60) propose a human–robot–system interaction model to predict and balance the ethical impact equally between not only caregivers and receivers, but for the care system within which these actors function. Along this line, Barrett *et al.* (61) analyze how introducing a pharmaceutical dispensing robot in a hospital changed the boundary dynamics of three occupational groups (pharmacists, technicians and assistants) and Mutlu & Forlizzi (62) found diametrically different responses in postpartum and cancer units to robots taking out laundry from patients’ rooms, making it clear that a holistic view is needed when planning the introduction of robots in any healthcare system or other workplaces.

At a macro level, 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 (55).

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 egalitarian access to services and, thus, the option of being redirected to a human agent should always be in place. This relates to human dignity as mentioned in Section 3.1.

Technology has a strong impact on the global distribution of wealth and power, causing divides per world areas. Nagenborg *et al.* (63) 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*».

Another form of potential discrimination towards certain collectives is related to robot appearance; it goes without saying that sexist, ableist, racist and ethnic robot morphologies and expressiveness in the design and programming of robots must be entirely avoided. But vulnerable minorities may suffer from less obvious bias-related concerns as they may not equally benefit from social robots enabled with certain technologies, such as face and expression recognition, since less exposure to their data may yield much higher error rates (64). By finding patterns within datasets that reflect implicit biases that permeate society, learning algorithms reinforce these biases. Howard & Borenstein (65) describe concrete examples of how bias has infused itself into current robotic systems, and how it may affect the future design of such systems.

On the positive side, 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 know-how to such groups (66).

4.2. Freedom

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 (66).

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 (67) refers to as «the watching eye and punitive hand of the state», which in an extreme way restrains 'negative freedom' or 'freedom from'. Berlin (68) distinguishes this type of freedom, namely the absence of constraints from others regarding one's own activity, from 'freedom to' or 'positive freedom', entailing the capacity to choose the reasons for one's choices.

The robot enforcement of certain habits and values on the user has been termed 'nudging' and discussed at the micro level in Section 3.2. This capacity may also lead to paternalism (69) which, at the macro level, infringes freedom by interfering in people's decisions even if presumably in their own interest. Thaler & Sunstein (35) 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).

4.3. Responsibility

In this section we address the responsibility of professionals and governors towards current society and future generations to ensure that robots are deployed to the maximum benefit of all citizens, and potential unintended consequences are proactively headed off (12). Most pressing global concerns are ecological sustainability of robots' life cycle, technological dependency, and ultimately the evolution of humankind with the potential loss of human capabilities if these are progressively delegated to robots.

Gunkel (70) discusses the two usual responses to the question of responsibility in social robotics, namely instrumentalism (the robot is a mere tool) and machine ethics (moral values and rules should be embedded in the robot itself). He advocates for hybrid responsibility to face the opportunity we have to collectively decide who to include in the community of moral subjects, and what we exclude from such consideration and why, which will have a profound effect on the way we conceptualize our place in the world. Assuming shared responsibility, Nyholm (71) analyses the types of agency that can and cannot be attributed to robotic systems, arguing that such agency should always be understood in terms of human–robot collaboration.

In the context of machine ethics, the issue of social responsibility underlies the inquiry discussed by Borenstein & Arkin (34): «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 robot design and deployment and, in turn, in the way of life of future generations. Symmetrically, humans have some responsibility as to how they treat robots; mainly due to the impact human-robot relationships have on human-human relationships, not only because of the isolation and the dehumanization of relationships they may cause (19), as mentioned in Section 3.4, but also because abusive behaviors towards robots may cause desensitization to these immoral behaviors at a societal level (72). This protection towards robots needs to be legally regulated.

Professional responsibility of roboticists include complying with ethics requirements in human-robot interaction research (73), as well as communicating properly (74). Regarding the former, Punchoojit & Hongwarittorrn (75) 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, privacy of their data must be guaranteed, and risks of physical or emotional harm should be minimized.

Moreover, Riek (76) provides detailed guidelines to ensure careful use of Wizard-of-Oz, a technique frequently employed by researchers in social robotics, 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. The possible fostering of inappropriate expectations among users must be taken into account, similarly as deceit caused by anthropomorphism was discussed in Section 3.3.

Turning to proper communication, Boden *et al.* (13) advises that «we, roboticists, should 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». Nourbakhsh (77) 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.

5. COURSE ON ETHICS OF SOCIAL ROBOTICS BASED ON SCIENCE FICTION

The eight issues described in Sections 3 and 4 follow from the following questions:

1. Could robot decision-making undermine human **dignity**?
2. Where is the boundary between helping and creating **dependency**?
3. Should the possibility of **deception** be actively excluded in the design of robots?
4. Is it acceptable for robots to behave as **emotional surrogates**?
5. When should a society's wellbeing and safety prevail over people's **privacy**?
6. What types of **discrimination** and digital divides may social robotics cause?
7. Could social robots be used to restrain **freedom** and control people?
8. How to deploy robots in a **responsible** way towards society and future generations?

These are among the 24 questions discussed in an ethics course, specifically focused on social robotics, which the author developed based on her novel *The Vestigial Heart* (78). The appeal of science fiction to debate about ethics has been widely recognized (79, 80), in particular by professors teaching this and similar courses (81): «Using fiction to teach ethics allows students to safely discuss and reason about difficult and emotionally charged issues without making the discussion personal.»

Instructors can download a teacher's guide and a 100-slide presentation free of charge from MIT Press website (78), as well as from the author's website (82). Rather than following a conceptual categorization as in the current paper, these teaching materials are organized with a more practical orientation centered on domains of application, such as healthcare, education, the workplace, and general social contexts, for which concrete examples are spelled out. Each section in the teacher's guide follows the same structure, starting with some highlighted scenes from a chapter in the novel, then the corresponding ethics academic background is provided, followed by four questions and hints for their discussion, and closing with some revisited issues from previous chapters. The transcription of the course, as delivered at a summer school, is also available (83).

For a discussion of the need for ethics education in technological university degrees and the current state of affairs in robotics curricula in particular, the reader is referred to (84, 4). Here we briefly mention that the trend towards increasing specialization that has dominated higher education in the last decades needs to be counterbalanced by adopting a wider view that takes into consideration the social implications of the technologies being studied. Computer science degrees have played a pioneering role in this regard. Renowned professors such as Barbara J. Grosz have long advocated for integrating ethics in computer science education: «By making ethical reasoning a central element in the curriculum, students can learn to think not only about what technology they could create, but also whether they should create that technology.» (85).

Along this line, prestigious associations such as the Association for Computing Machinery (ACM), the Institute of Electrical and Electronics Engineers (IEEE) and the Association for the Advancement of Artificial Intelligence (AAAI) include 17 knowledge areas in their Computer Science curricula (86), one of which is Society, Ethics and Professionalism, so that «students develop an understanding of the relevant social, ethical, legal and professional issues». The impact of AI on individual wellbeing and civic life is prominently featured in the above curricula and recognized as a crosscutting theme. Courses on AI Ethics have quickly proliferated, most of them being taught online. Some include one or two sessions on the ethics of human-robot interaction, but there are very few entire courses on Robot Ethics, and less so on the Ethics of Social Robotics.

The envisaged expansion of social robots will lead to a growing need for practical and engaging ethics courses devoted to them, not only in Computer science, Electrical engineering and Philosophy curricula, but also in Anthropology, Psychology, Sociology, Political science, Economics, Business administration, and related disciplines in the Social sciences and the Humanities, given the multidisciplinary nature of the issues involved. The embodiment and cost of social robots makes their deployment to occur at a much slower pace than that of AI applications, providing an opportunity to agree on a basic set of criteria and ethical principles to be taught. The IEEE initiative (12) is an attempt in this direction, and some courses are starting to appear based on it (87). The present review has the same aim, but focused on the specificities of social robots.

6. OPPORTUNITIES AND FUTURE PROSPECTS

If initially robot ethics was mainly tackled by roboticists, in the last years philosophers and social scientists have joined them in multidisciplinary teams. This has widened the angle of the ethics lens, extending previous reflection along two lines: i) considering more global societal aspects beyond dyadic human-robot interaction, as already discussed in Section 4, and ii) viewing the deployment of social robots as providing unprecedented opportunities for new forms of relationship that favor human self-knowledge, personal growth, and the development of social values and cohesion.

Along this second line of reflection, Coeckelbergh (88, 89) advocates for turning to a philosophy of interaction in order to establish an ethics of appearance and human good. 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. 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. In order to explore alternative robotic morphologies that could enrich people's daily interactions, Sirkin & Ju (90) have robotized some everyday objects to appropriately respond to human intentions and emotions, and Sabanovic *et al.* (91) have proposed innovative prototyping methods for designing socially situated embodiments.

A related proposal is that of synthetic ethics, which claims that ethics develops as a joint learning process through active participation in the conception, development and carrying out of projects, rather than judging from outside the results of ongoing research and application. Dumouchel & Damiano (92) promote this approach to gain a better understanding of ourselves, especially the affective and social aspects of our minds. Rajaonah & Zio (93) rely also on synthetic ethics by proposing a methodology to analyze the co-construction of ethical interactions between humans and social robots, whereby ethical know-how is gained on both sides.

Abounding on the opportunity robots offer to reflect on ourselves as human beings, Broadbent (94) outlines the value and utility of using robots to examine a number of fundamental features of human behavior, perception and cognition, a work that has been very influential for researchers working at the intersection of cognitive neuroscience and HRI (95, 96).

Another collaborative approach to the development of robot ethics stems from the need to bridge the gap between philosophical and empirical research in each particular application domain. For instance, in the use of social robots for elderly care, Vandemeulebroucke *et al.* (97) highlight the importance of grounding philosophical-ethical reflection on the empirical-ethical knowledge of older adults and their caregivers. They propose a framework for this philosophical-empirical dialog that opens the ethics of assistive robotics for the elderly to its own socio-historical contextualisation.

As devices that mediate actions (98), robots not only transform the practices carried out in an environment, but also its characteristic values. Given the profoundly transformative potential of introducing social robots, particularly in care-giving and educational environments, Toboso *et al*. (99) propose to study its effects on the wellbeing and quality of life of people and communities by means of Nussbaum's capabilities approach (100). This is a powerful proposal of ways in which societies can promote justice through encouraging the development of capabilities that are essential to what it means to be human.

The development or transformation of individual and social values by the action of robots can be intentional or unintentional. As an example of the former in a children's playground, 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.

To illustrate how robots could promote social justice, Borenstein & Arkin (34) 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.

Besides social values, another human trait that is highly valued is creativity. As anthropologists and economists know well, people can be very creative in their use of objects, turning them away from their original design. This capability could be enhanced through new ways of relating to interacting objects, as mentioned earlier.

To summarize, the ethics of social robotics goes well beyond the regulation of robot behavior, and it aims to understand the characteristics of new types of social artificial agents and to discover new forms of interaction between humans and different such agents. Thus, it is a dynamic endeavor that needs to learn from and deal with changing demands, contexts and constraints.

As the above glimpse through the wide-angle ethics lens suggests, social robotics opens amazing future perspectives. More so, because the field is developing at a pace that still permits to intertwine ethics in research and deployment. However, this requires extensive ethics education at all levels, particularly in related technological university degrees (84), as well as researchers to take the ethical implications of our work seriously, which needs to be reflected in our scientific papers and outreach activities for society at large to develop trust towards social robots. This is our responsibility towards current society and future generations.

ACKNOWLEDGEMENT

This work has been partly supported by the European Research Council (ERC) through the project CLOTHILDE (Advanced Grant agreement No 741930), and the European Union's Horizon Europe Programme through the project SoftEnable (HORIZON-CL4-2021-DIGITAL-EMERGING-01-101070600).

REFERENCES

1. Torras C. 2016. Service robots for citizens of the future. *Eur. Rev.* 24:17-30

2. Pareto J, Román B, Torras C. 2021. The ethical issues of social assistive robotics: A critical literature review. *Technol. Soc.* 67:101726

3. Simon HA. 1969. *The Sciences of the Artificial*. Cambridge, MA: MIT Press

4. Torras C. 2020. Science-fiction: A mirror for the future of humankind. *IDEES* 48:1-11

5. Heidegger M. 1954. Die Frage nach der Technik (The Question Concerning Technology). In *Vorträge und Aufsätze*. Stuttgart: Verlag Günther Neske

6. Jonas H. 1979. *Das Prinzip Verantwortung: Versuch einer Ethik für die technologische Zivilisation* (*The Imperative of Responsibility: In Search of an Ethics for the Technological Age)*. Frankfurt: Insel-Verlag

7. Bilbeny N. ed. 2023. *Robótica, ética y política (Robotics, ethics and politics)*. Barcelona: Icaria Editorial

8. Carbonell E. 2022. *El futur de la humanitat (The future of humankind)*. Barcelona: Ara Llibres

9. Veruggio G, Solis J, Van der Loos M. 2011. Roboethics: Ethics applied to robotics. *IEEE Robot. Autom. Mag.* 18:21-22

10. Sullins JP. 2015. Applied professional ethics for the reluctant roboticist. In *Proc. 10th ACM/IEEE International Conference on Human-Robot Interaction: The Emerging Policy and Ethics of Human-Robot Interaction Workshop*, pp. 1-8. Piscataway, NJ: IEEE

11. Wallach W, Allen C. 2008. *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press

12. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. 2018. *Ethically aligned design: A vision for prioritizing human wellbeing with autonomous and intelligent systems*, Version 2. Piscataway, NJ: IEEE. https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf

13. 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, Winfield AF. 2017. Principles of robotics: Regulating robots in the real world. *Connect. Sci.* 29:124-29

14. European Parliament. 2017. *Civil law rules on robotics.* https://www.europarl.europa.eu/doceo/document/TA-8-2017-0051_EN.html

15. EU High-level Expert Group on AI. 2019. *Ethics guidelines for trustworthy AI.* https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419

16. Jobin A, Ienca M, Vayena E. 2019. The global landscape of AI ethics guidelines. *Nat. Mach Intell.* 1:389-99

17. Sabater A, de Manuel A. 2022. *The PIO model (Principles, Indicators and Observables): A proposal for organizational self-assessment on the ethical use of data and artificial intelligence systems*. OEIAC: Observatori d'Ètica en Intel·ligència Artificial de Catalunya. https://www.udg.edu/ca/Portals/57/OContent_Docs/modelpio_ENG_v4.pdf

18. Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, Luetge C, Madelin R, Pagallo U, Rossi F, Schafer B. 2018. AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds Mach.* 28:689-707

19. Fosch-Villaronga E, Lutz C, Tamò-Larrieux A. 2020. Gathering expert opinions for social robots' ethical, legal, and societal concerns: Findings from four international workshops. *Int. J. Soc. Robot.* 12:441-58

20. Vandemeulebroucke T, de Casterlé BD, Gastmans C. 2018. The use of care robots in aged care: A systematic review of argument-based ethics literature. *Arch. Gerontol. Geriatr.* 74:15-25.

21. RoboLaw Project. 2014. Deliverable D6.2 - *Guidelines on Regulating Robotics*. http://www.robolaw.eu/RoboLaw_files/documents/robolaw_d6.2_guidelinesregulatingrobotics_20140922.pdf

22. Feil-Seifer D, Mataric MJ. 2011. Socially assistive robotics - ethical issues related to technology. *IEEE Robot. Autom. Mag. 18*:24-31

23. Sharkey A, Sharkey N. 2012. Granny and the robots: ethical issues in robot care for the elderly. *Ethics Inf. Technol.* 14:27-40

24. Lichocki P, Kahn Jr PH, Billard A. 2011. A survey of the robotics ethical landscape. *IEEE Robot. Autom. Mag.* 18:39-50

25. Cowie R. 2015. Ethical issues in affective computing. In *The Oxford Handbook of Affective Computing*, pp. 334-348

26. Sharkey A. 2014. Robots and human dignity: a consideration of the effects of robot care on the dignity of older people. *Ethics Inf. Technol. 16*:63-75

27. Sharkey N, Sharkey A. 2010. The crying shame of robot nannies: An ethical appraisal. *Interact. Stud.* 11:161-90

28. Wilks Y. 2010. Introducing artificial companions. In *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues*, ed. Y Wilks, Natural Language Processing Series 8. Amsterdam: John Benjamins Publishing.

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

30. Espingardeiro A. 2015. Social assistive robots, reframing the human robotics interaction benchmark of social success. *Int. J. Soc. Behav. Educ. Econ. Bus. Ind. Eng.* 9:377-82

31. 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

32. Roberts R. 2001. *The Invisible Heart – An Economic Romance*. Cambridge, MA: MIT Press

33. Castellano G, De Carolis B, D'Errico F, Macchiarulo N, Rossano V. 2021. PeppeRecycle: Improving children's attitude toward recycling by playing with a social robot. *Int. J. Soc. Rob.* 13:97-111

34. Borenstein J, Arkin R. 2016. Robotic nudges: The ethics of engineering a more socially just human being. *Sci. Eng. Ethics* 22:31-46

35. Thaler RH, Sunstein CR. 2008. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. New Haven: Yale University Press

36. Li J. 2013. The nature of the bots: how people respond to robots, virtual agents and humans as multimodal stimuli. In *15th ACM International Conference on Multimodal Interaction*, pp. 337–40

37. Winfield AF, Booth S, Dennis L, Egawa T, Hastie H, Jacobs N, Muttram R, Olszewska J, Rajabiyazdi F, Theodorou A, Underwood M, Wortham RH, Watson E. 2021. IEEE P7001: A proposed standard on transparency. *Front. Robot. AI* 8:665729

38. Sharkey A, Sharkey N. 2021. We need to talk about deception in social robotics! *Ethics Inform. Technol.* 23:309-16

39. Van Maris A, Zook N, Caleb-Solly P, Studley M, Winfield A, Dogramadzi S. 2018. Ethical considerations of (contextually) affective robot behaviour. In *Hybrid Worlds: Societal and Ethical Challenges*, ed. S Bringsjord, MO Tokhi, MIA Ferreira, NS Govindarajulu, pp. 13-19

40. Matthias A. 2015. Robot lies in health care: When Is deception morally permissible? *Kennedy Inst. Ethics J.* 25:169-192

41. Guizzo E. 2015. The little robot that could... Maybe. *IEEE Spectr*. 53:58-62.

42. Riek LD, Howard D. 2014. A code of ethics for the human-robot interaction profession. In *We Robot Conference,* pp. 1-10

43.Van der Loos HM. 2007. Ethics by design: A conceptual approach to personal and service robot systems. In *Proc. Roboethics Workshop at IEEE International Conference on Robotics and Automation.* Piscataway, NJ: IEEE

44.Dautenhahn K, Woods S, Kaouri C, Walters ML, Koay KL, Werry I. 2005. What is a robot companion-friend, assistant or butler? In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1192-1197. Piscataway, NJ: IEEE

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

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

47.Bryson JJ. 2010. Robots should be slaves. In *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues*, ed. Y Wilks, pp. 63-74. Amsterdam: John Benjamins Publishing.

48.Sullins J. 2012 Robots, love, and sex: the ethics of building a love machine. *IEEE Trans. Affect. Comput.* 3:398–409

49.de Graaf MM. 2016. An ethical evaluation of human–robot relationships. *Int. J. Soc. Robot. 8*:589-98

50.Vallverdú J, Casacuberta D. 2014. Ethical and technical aspects of emotions to create empathy in medical machines. In *Machine Medical Ethics*, ed. SP van Rysewyk, M Pontier, pp. 341-62. Cham: Springer International Publishing

51.Turkle S. 2007. Authenticity in the age of digital companions. *Interact. Stud. 8*:501-17

52.Riek LD, Rabinowitch TC, Chakrabarti B, Robinson P. 2009 Empathizing with robots: Fellow feeling along the anthropomorphic spectrum. In *3rd IEEE Intl. Conf. on Affective Computing and Intelligent Interaction and Workshops*, pp. 1-6

53.Reig S, Carter EJ, Tan XZ, Steinfeld A, Forlizzi J. 2021. Perceptions of agent loyalty with ancillary users. Int. J. Soc. Robot. 13:2039-55

54.Calo R. 2015. Robotics and the lessons of cyberlaw. *Calif. Law Rev.* 103:513-63

55. Veruggio G, Operto F, Bekey G. 2016. Roboethics: Social and ethical implications of robotics. In *Springer Handbook of Robotics*, *2nd edition*, ed. B Siciliano, O. Khatib, Chapter 80, pp. 2135-60. Cham, Switz.: Springer

56. Verbeek P. 2008. Morality in design: Design ethics and the morality of technological artifacts. In *Philosophy and Design*, ed. PE Vermaas, P Kroes, A Light, SA Moore S.A, pp. 91-103. Cham, Switz.: Springer

57. Bonnemains V, Saurel C, Tessier C. 2018. Embedded ethics: some technical and ethical challenges. *Ethics Inf. Technol*. *20*:41-58.

58. Wallach W. 2010. Robot Morals and Human Ethics: The Seminar, *Teach. ethics* 11:87-92

59. Leroux C, Labruto R. 2012. Ethical, legal, and societal issues in robotics. *euRobotics: The European Robotics Coordination Action*, Deliverable D3.2.1.

60. van Wynsberghe A, Li S. 2019. A paradigm shift for robot ethics: from HRI to human–robot–system interaction (HRSI). *Medicolegal Bioeth.* 9, 11–21.

61. Barrett M, Oborn E, Orlikowski WJ, Yates J. 2012. Reconfiguring boundary relations: Robotic innovations in pharmacy work. *Organ. Sci.* 23:1448-66

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

63. Nagenborg M, Capurro R, Weber J, Pingel C. 2008. Ethical regulations on robotics in Europe. *AI Soc.* 22:349-366

64. Buolamwini J, Gebru T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77-91

65. Howard A, Borenstein J. 2018. The ugly truth about ourselves and our robot creations: The problem of bias and social inequity. *Sci, Eng. Ethics* 24:1521–36

66. Peltu M, Wilks Y. 2010. Summary and discussion of the issues. In *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues*, ed. Y Wilks, pp. 259-86. Amsterdam: John Benjamins Publishing.

67. Lowe W. 2010. Identifying your accompanist. In *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues*, ed. Y Wilks, pp. 95-100. Amsterdam: John Benjamins Publishing.

68. Berlin I. 1969. *Four Essays on Liberty*. Oxford University Press

69. Coeckelbergh M. 2022. The Political Philosophy of AI: An Introduction. Cambridge, UK: Polity Press

70. Gunkel DJ. 2020. Mind the gap: responsible robotics and the problem of responsibility. *Ethics Inform. Technol.* 22:307-20

71. Nyholm S. 2017. Attributing agency to automated systems: Reflections on human–robot collaborations and responsibility-loci. *Sci. Eng. Ethics* 24*:*1–19

72. Darling K. 2016. Extending legal protection to social robots: The effects of anthropomorphism, empathy, and violent behavior towards robotic objects. In *Robot law*, ed. R Calo, AM Froomkin, I Kerrpp, pp. 213-32. Cheltenham: Edward Elgar Publishing

73. Rosén J, Lindblom J, Billing E, Lamb, M. 2021. Ethical challenges in the human-robot Interaction field. In *TRAITS Workshop at ACM/IEEE International Conference on Human-Robot Interaction*, ACM Digital Library

74. Laumond JP, Danblon E, Pieters C. ed. 2019. Wording Robotics: Discourses and Representations on Robotics. *Springer Tracts in Advanced Robotics* 130. Springer.

75. Punchoojit L, Hongwarittorrn N. 2015. Research ethics in human-computer interaction: A review of ethical concerns in the past five years. In *2nd National Foundation for Science and Technology Development Conference on Information and Computer Science*, pp. 180-85

76. Riek LD. 2012. Wizard of Oz studies in HRI: A systematic review and new reporting guidelines. *J. hum. robot interact.* 1:119-36

77. Nourbakhsh I. 2010. The rhetorics of robotics. Unpublished manuscript. https://sites.google.com/site/ethicsandrobotics/ethics-and-robotics-a-teaching-guide/living-archive/reading/nourbakhsh

78. Torras C. 2018. *The Vestigial Heart. A Novel of the Robot Age*. Cambridge, MA: MIT Press (Teacher's guide and 100-slide presentation freely available to instructors at https://mitpress.ublish.com/book/the-vestigial-heart-a-novel-of-the-robot-age#ancillaries)

79. Torras C. 2010. Robbie, the pioneer robot nanny: Science fiction helps develop ethical social opinion. *Interact. Stud.* 11:269-73

80. Torras C. 2023. La ciencia ficción como estímulo del debate ético en robótica (Science fiction as a stimulus for ethics debate in robotics). In *Robótica, ética y política (Robotics, ethics and politics)*, ed. N Bilbeny, pp. 139-67. Barcelona: Icaria Editorial

81. Burton E, Goldsmith J, Mattei N. 2018. How to teach computer ethics through science fiction. *Commun. ACM.* 61:54-64

82. http://www.iri.upc.edu/people/torras/vestigial.html

83. Torras C, Ludescher LG. 2023. Writing science fiction as an inspiration for AI research and ethics dissemination. In *Human-Centered Artificial Intelligence*, ed. M Chetouani, V Dignum, P Lukowicz, C Sierra, *Lect. Notes Comput. Sci*. 13500:322-44. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-24349-3_17

84. Torras C. 2019. Robotics and artificial intelligence meet the humanities: Some initiatives for ethics education and dissemination. In *Humanities and Higher Education: Synergies between Science, Technology and Humanities*, pp. 267-273. Barcelona: Global University Network for Innovation

85. Grosz BJ, Grant DG, Vredenburgh K, Behrends J, Hu L, Simmons A, Waldo J. 2019. Embedded EthiCS: integrating ethics across CS education. *Commun. ACM* 62:54-61

86. ACM/IEEE-CS/AAAI Computer Science Curricula. 2023. https://csed.acm.org/

87. The Institution of Engineering and Technology (IET). https://academy.theiet.org/robot-ethics

88. Coeckelbergh M. 2009. Personal robots, appearance, and human good: A methodological reflection on roboethics*. Intl. J. of Soc. Robot.* 1:217-21

89. Coeckelbergh M. 2022. Three responses to anthropomorphism in social robotics: Towards a critical, relational, and hermeneutic approach. *Intl. J. of Soc. Robot.* 14:2049-61

90. Sirkin D, Ju W. 2014. Using embodied design improvisation as a design research tool. In *International Conference on Human Behavior in Design*, pp. 14-17

91. Sabanovic S, Reeder S, Kechavarzi B. 2014. Designing robots in the wild: In situ prototype evaluation for a break management robot. *J. hum. robot interact. 3*:70-88

92. Dumouchel P, Damiano L. 2017. *Living with robots*. Cambridge, MA: Harvard Univ. Press.

93. Rajaonah B, Zio E. 2022. Social robotics and synthetic ethics: A methodological proposal for research. *Int. J. Soc. Rob.* https://doi.org/10.1007/s12369-022-00874-1

94. Broadbent E. 2017. Interactions with robots: the truths we reveal about ourselves. *Annu. Rev. Psychol.* 68:627-652

95. Pérez-Osorio J, De Tommaso D, Baykara E, Wykowska A. 2018. Joint action with Icub: a successful adaptation of a paradigm of cognitive neuroscience in HRI. In *27th IEEE International Symposium on Robot and Human Interactive* Communication, pp. 152–57

96. Henschel A, Laban G, Cross ES. 2021. What makes a robot social? A review of social robots from science fiction to a home or hospital near you. *Curr. Robot. Rep. 2*:9-19

97. Vandemeulebroucke T, Casterle BD, Gastmans C. 2020. Ethics of socially assistive robots in aged-care settings: A socio-historical contextualisation. *J. Med. Ethics* 46:128–36

98. Verbeek PP. 2015. Beyond interaction: a short introduction to mediation theory. *Interact.* 22:26-31

99. Toboso M, Morte R, Monasterio A, Ausín T, Aparicio M, López D. 2020. Robotics as an instrument for social mediation. In *Inclusive Robotics for a Better Society: Selected Papers from INBOTS Conference 2018,* pp. 51-58. Cham: Springer International Publishing

100. Nussbaum MC. 2011. *Creating Capabilities: The Human Development Approach*. Cambridge, MA: Harvard University Press