FEATURES

24 Socially Assistive Robotics
Ethical Issues Related to Technology
By David Feil-Seifer and Maja J. Matarić

32 Children, the Elderly, and Interactive Robots
Anthropomorphism and Deception in Robot Care and Companionship
By Amanda Sharkey and Noel Sharkey

39 The Ethical Landscape of Robotics
Bringing Ethics into the Design and Use of Robots
By Pawel Lichocki, Peter H. Kahn, Jr., and Aude Billard

51 Incremental Machine Ethics
Adaptation of Programmed Constraints
By Thomas M. Powers

59 The Robot DustCart
Focus on Social and Legal Challenges
By Pericle Salvini, Giancarlo Teti, Enza Spadoni, Cecilia Laschi, Barbara Mazzolai, and Paolo Dario

68 Remote-Control Crimes
Roboethics and Legal Jurisdictions of Tele-Agency
By Peter M. Asaro

72 Ethics in Advanced Robotics
ELS Issues in Advanced Robotics
By Fiorella Operto

79 Motion Planning
Part I: The Essentials
By Steven M. LaValle
Roboethics: Ethics Applied to Robotics

By Gianmarco Veruggio, Jorge Solis, and Machiel Van der Loos

This special issue deals with the emerging debate on roboethics, the human ethics applied to robotics. Is a specific ethic applied to robotics truly necessary? Or, conversely, are not the general principles of ethics adequate to answer many of the issues raised by our field’s applications? In our opinion, and according to many roboticists and human scientists, many novel issues that emerge and many more that will show up in the immediate future, arising from the upcoming marketed robotics products, demand the development of new cultural and legal tools that can provide the crucial answers to the most sensitive questions.

The unfolding and emerging scenarios made possible by robotics are fascinating and unsettling at the same time. Suffice it to think that all machines, of any form and dimension and for any type of use, will be computerized, equipped with artificial intelligence and networked, to understand that everything we have seen to date—computers, video games, cellular phones, and Internet—is really only the dawn of the technological world that awaits us. For instance, in aging societies, there is an urgent motivation for safe, autonomous, and adaptable personal (also called social) robots. So humans will coexist with the next-generation robots employed as domestic workers, nurses, and caregivers at home, in hospitals, and in nursing homes.

This widespread distribution of robots will raise several completely new ethical, legal, and social issues. Robots will have the ability to learn and process our personal profiles, tastes, and habits, which will lead to privacy and safety issues, as well as those regarding individual freedom. The human–robot interactions can cause psychological and social problems, especially in vulnerable populations such as children, older persons, and patients. Then there will be issues regarding the attribution of civil and criminal liability should an autonomous robot produce damages. Finally, there will be important, critical areas bordering with bioethics, in cases of medical and biorobotics, and with humanitarian and international law, in cases of military robotics. All these cases have never been faced squarely by humanity, and this entails a need for a complex, joint approach from various disciplines to handle them.

These issues have been subject to discussion since the dawn of robotics in the works of Norbert Wiener or in the science-fiction speculations of Isaac Asimov. However, it is only in the last few years that the debate has been progressively organized within the international robotics community and that the key word roboethics has established itself as an emerging field of applied ethics. The complexity of the matter is enormous, as is the tableau painted by the various overlapping scientific and cultural backgrounds in the debate. This is why we believe it is worth addressing the terminology issue in this introduction to clarify the interconnecting levels between ethics and robotics.

The first level is represented by the adopted ethical theories, developed principally by the branch of philosophy called ethics or morality, which studies human conduct, moral assessments, and the concepts of good and evil, right and wrong, justice and injustice, and so on. In our case, a generic or fundamental ethical reflection is directly related to the particular issues that are generated by the development of robotic applications and their diffusion in the society. This is the proper concept of roboethics, meaning that applied ethics, similar to bioethics, attempts to provide answers to new questions that are generated by the progress of a specific scientific and technical field.

This level updates the various views on concepts, such as dignity and integrity of the person and the fundamental rights of the individual, as well as the social, psychological, and legal aspects involved. The second level, currently referred to as robot ethics or machine ethics, regards the code of conduct that designers implement in the artificial intelligence of robots. This means a sort of artificial ethics able to guarantee that autonomous robots will exhibit ethically acceptable behavior in all situations in which they interact with human beings or when their actions may have negative consequences on human beings or the environment. It is clear that the guidelines to define what is ethically acceptable and to enforce them are the product of the aforementioned field of roboethics. Robots are, in fact, machines, meaning tools that are unaware of the...
choices made by their human creators, which, therefore, bear the moral responsibility for the actions, good or bad, of robots.

Finally, there is a third level, which we could perhaps define as robot’s ethics, because it is the ethic born from the subjective morality of a hypothetical robot that is equipped with a conscience and freedom to choose its own actions on the basis of a full comprehension of their implications and consequences. It is only in this case that robots may be deemed as moral agents and that one may refer to as involving the responsibilities or rights of robots. This, obviously, is currently speculative and beyond the purposes of this special issue.

It is, therefore, clear that roboethics is a work in progress, susceptible to further evolution as the events unroll in our technical and scientific future. We are convinced that all stakeholders in the development of robotics must take part, starting with the robotics scientists and also all members of the Society. The role of the media will be crucial to this: they will have to provide prompt and correct information on the progress of robotics and the pros and cons of its applications. An even more important role will be played by the world’s school systems, which will have the task of training upcoming generations: the true players, beneficiaries or victims, of the imminent robotics invasion.

This special issue, being the first dedicated to the topic of roboethics and given the high number of submissions and the limited available space, gives priority to broader articles that provide cultural and philosophical direction to those approaching the subject for the first time and will publish some articles analyzing the human–robot relationship from various points of view: technical, psychological, sociological, and legal. Other sensitive topics, such as military robotics or birobotics, will require further and deeper ethical analysis in future issues of the magazine. In the following paragraphs, we briefly discuss the content of each article.

The first article, “Socially Assistive Robotics,” by Feil-Seifer and Mataric, examines the ethical issues involved in using socially assistive robots, particularly in the context of health care. They describe core ethical principles for robots that provide assistance through social interaction, and they emphasize how deception (intended or unintended), autonomy, and justice can affect the ethical applications of assistive robots.

The topic is further investigated in “Children, the Elderly, and Interactive Robots” by Sharkey and Sharkey, who examine the complex psychological implications of the relationships with robots, mainly through theoretical references to cognitive psychology. They start from a survey of the present state of the art in robot caregivers and pets and discuss the risks and benefits of the relational applications with the oldest and youngest members of Society.

In “The Ethical Landscape of Robotics,” Lichocki et al. survey some of the main ethical issues pertaining to robotics that have been discussed in the literature so far. They start with the notion of responsibility ascription that arises when an autonomous system malfunctions or harms people. Then, they list various ethical issues emerging in two sets of robotic applications: service robots that peacefully interact with humans and lethal robots created to fight in the battlefields. Finally, they also provide a short overview of machine ethics.

Powers broadens the ongoing debate on machine ethics, adding an incremental strategy. In his approach, incrementalism in machine ethics becomes a practical proposal about how to simultaneously engineer and provide ethical sanction for robots. The article discusses the concrete proposals to do this and reflects in a critical manner on these matters.

A very interesting experimental approach is that described by Salvini et al. in “The Robot DustCart.” The article describes DustCart, a project concerning the use of autonomous mobile robots to collect and transport rubbish bags in a small Italian town. After a report on the testing period (service provided, testing site, and so on), the authors deal with the social and legal implications of the experiment.

A further reflection on legal aspects is given in Asaro’s article, “Remote-Control Crimes,” which deals with the difficult international and cross-cultural aspects of roboethics. He discusses the difficulties of applying law to criminal activities that will be enabled in the future by new robotic capabilities, such as cybercrimes; robot crimes will be the subject of multiple governing laws, changing national rules, conflicting regulations, and disparate institutions.

Finally, in “Ethics in Advanced Robotics,” Operto outlines a brief history of roboethics, whose development she has contributed to since its birth; in her article, she points out the need to uncover the philosophical assumptions underlying today’s debate in ethical and social issues of robotics to facilitate the establishment of a common ground for the definition of principles and regulatory guidelines.

We hope that the readers will enjoy the articles in this special issue, are encouraged to deepen their interest in roboethics, and will actively contribute to the debate, which will become increasingly important with the growth of robotics in the society.
Socially Assistive Robotics

Ethical Issues Related to Technology

By David Feil-Seifer and Maja J. Matarić
Socially assistive robotics (SAR) aims to address critical areas and gaps in care by automating supervision, coaching, motivation, and companionship aspects of one-on-one interactions with individuals from various large and growing populations, including stroke survivors, the elderly and individuals with dementia, and children with autism spectrum disorders (ASDs). This article examines the ethical challenges of SAR from three points of view (user, caregiver, and peer) using core principles from medical ethics (autonomy, beneficence, nonmaleficence, and justice) to determine how intended and unintended effects of SAR can impact the delivery of care.

**Socially Assistive Robotics**
The most obvious and direct risk of any assistive technology, including SAR, is the potential of physical harm. While this is an important risk to examine, SAR is primarily concerned with robots that provide assistance through social, rather than physical, interaction. In this article, we outline the commonly accepted core principles from medical ethics and use those principles as guidelines for evaluating the risks of SAR. We use examples of SAR systems to describe the ways that robots are currently being used as directions for future use based on an ongoing research. We then discuss the core ethical principles to be examined. Finally, we apply each principle to SAR in turn and discuss its implications.

**Definition of SAR**
SAR [5] describes a class of robots that is the intersection of assistive robotics (robots that provide assistance to a user) and socially interactive robotics (robots that communicate with a user through social and nonphysical interaction). Assistive robotics is a broad class of robots whose function is to provide assistance to users, ranging from getting out of bed, brushing teeth, locomotion, and rehabilitation. This section provides few examples of SAR systems.

Wada et al. [23] describes the design of Paro, a robot for pet-therapy applications for nursing homes that do not allow pets. Pet therapy has been shown to have a positive effect on the elderly in nursing-home settings [16], but there are logistical challenges to having animals in nursing homes. Paro was built to resemble a baby harp seal and designed to interact like a pet with simple sounds and movements made in response to being held and petted. Experimental results suggest that Paro may be effective for reducing stress in nursing-home residents. In addition, when placed in common areas of nursing homes, it produced increased social activity among residents. This suggests that SAR systems may be useful not just for their direct therapeutic applications but more generally as catalysts for social interaction.

Another SAR system is Roball [18], a self-propelling robotic ball that can sense its position and motion and thus the way it is being played with. Roball is being evaluated for use by children, including children with ASDs in the home or in clinical settings. Children with ASD typically have decreased social interactive behavior; encouraging play with therapists, family, and peers could have both diagnostic and therapeutic uses. Roball and other robots for play could be used as an addition to current ASD diagnostics or therapeutic regimens or as tools for developing new diagnostic and...
therapy methods. In general, the aim of SAR for ASD is to encourage children to initiate and sustain social interaction [17] with a parent, therapist, sibling, or peer.

Poststroke rehabilitation is another area where SAR can provide therapeutic benefits. Rehabilitation robotics has been developing robot arms that apply and measure forces on the user’s limbs. Such hands-on movement training is particularly useful in the early stages poststroke. However, a major long-term challenge of poststroke recovery, and rehabilitation in general, is encouraging compliance with the prescribed therapeutic regimen. Mataric et al. [12] describes a SAR system designed to improve therapeutic compliance through verbal noncontact coaching and encouragement. Such systems are designed to work in concert with the established stroke exercise methods such as constraint-induced therapy, building on and augmenting effective health-care practices.

Concurrent with the developing SAR technologies, ethical appraisal studies are being conducted about their acceptance. Mutlu and Forlizzi [4] conducted an ethnographic study of a delivery robot used in multiple departments of a hospital, finding that different patient groups had different reactions to the robot. For example, cancer units were not accepting the robot, finding it annoying, while postpartum units were accepting the robot and calling it delightful. The results of this study suggest that user populations could have completely different experiences with the same robot and that these experiences could be based on the users’ preexisting social and task dynamics and context. Tapus et al. [21] described a study in which elderly participants with Alzheimer’s disease interacted with a SAR robot that promoted cognitive exercises through a song-recognition game in a six-month study. The study participants included the robots in their narratives and preferred it to a computer. Turkle [22] demonstrated that some participants interacting with robots can correctly identify the robot’s intended emotional abilities and operational capabilities. These participants could also correctly distinguish equivalent capabilities in a person, pet, or other relational artifact. However, it was also demonstrated that some users formed attachments and emotional bonds with the robots they were interacting with. These attachments led to misconceptions about the robots’ emotional capabilities. For example, one user felt that the robot would miss him when he was gone, which is something that the robot was not capable of doing. In their hyperbolic yet poignant article, Sharkey and Sharkey [19] argue that such attachments in children could lead to malformed development and emotional problems.

**Persons Affected by SAR**

SAR is designed for use in a wide variety of settings including hospitals, schools, elder-care facilities, and private homes. The intended end users of such systems are individuals with special needs, but SAR systems must operate in real-world environments that may also include family, caregivers, and medical personnel. Consequently, the effects of SAR must be assessed for all of the individuals affected by the technology.

**Core Ethical Principles**

There are many ways to approach potential ethical issues related to technology in general, and SAR in particular. Several appraisals of specific SAR systems have been implemented and some have discussed the ethical dilemmas that a particular system poses [19], [22]. Studies have also aimed to establish ethical benchmarks related to the design, manufacture, or use of SAR [7], [9]. Finally, some appraisals have applied the core ethical principles to identify potential problems [3]. In this work, we apply an established medical ethics framework to identify potential issues related to SAR. This framework uses the following core principles for considering ethical issues [4]:

- **beneficence**: caregivers should act in the best interest of the patient
- **nonmaleficence**: the doctrine, “first, do no harm,” followed by the caregivers to avoid harming patients
- **autonomy**: the capacity to make an informed, uncoerced decision about care
- **justice**: fair distribution of scarce health resources.

There is dissension about whether or not the Beauchamp and Childress model is the ideal model for assessing medical ethics, as the foundation for current ethical appraisal and ethical training, we feel it is a sufficient jump-off point for discussion.

These principles underlie the ethical reviews of experiments with human participants and can also thus provide broad categories for examining ethical issues related to SAR. To perform such an examination, we use examples from actual SAR system experiments. However, these descriptions are only considerations of hypothetical scenarios and not meant to make judgments on the ethical validity of those specific SAR systems. In the next section, we describe the principles of beneficence and nonmaleficence and how they relate to the ethical use of SAR.

**Beneficence and Nonmaleficence**

The principles of beneficence and nonmaleficence state that caregivers should act in the best interests of the patient and should do nothing rather than take any action that may harm a patient. These principles establish that the potential benefits of an ethical treatment should exceed the risks. SAR, like any technology, features some risks along with the compelling potential benefits.

As noted earlier, SAR technologies are typically noncontact, so physical risk, while usually the most obvious ethical
Concern, is not a major issue of concern. SAR systems are designed so the robot does not apply any forces on the user. On the other hand, the user can touch the SAR system, and in some cases (as with Paro, see earlier), such contact is part of the therapy. However, in a majority of systems no physical contact is involved, and the robot may not even be within reach of the user, though it is typically within the social interactive space conducive to one-on-one interaction through speech, gesture, and body movement.

In this section, we examine some of the aspects of SAR technologies that are unique and ways in which SAR systems, in particular, might impact not only the user directly but also others in the shared context. In particular, the most prominent nonphysical risks posed by SAR systems include, but are not limited to, attachment to the robot, deception about the abilities of the robot, and influence on the human–human interaction of a robot’s user.

**Relationships, Authority, and Attachment**

It is safe to assume that a robot would not be the only caregiver/therapist for an assisted individual. Typically, care is provided by human caregivers, including professionals and family members. Thus, the SAR system impacts all of these individuals in various ways. For example, a robot that does something that a human caregiver would otherwise do (e.g., providing encouragement for performing exercises) might have as much impact on the human caregiver as on the patient, through the reduction of tasks related to a patient or through the reduction of workplace monotony. Specifically, many SAR systems are being designed to reduce the burden and burnout of family members and other caregivers. A SAR system might also provide a benefit to a caregiver by monitoring multiple aspects of the patient and providing ongoing quantitative assessments.

Sharkey and Sharkey [19] described another significant ethical dilemma that occurs when a user becomes emotionally attached to the robot. While establishing engagement and having the user enjoy interactions with the robot is a goal of SAR, attachment can also result in problems under certain circumstances. For example, if the robot’s effectiveness wanes, its scheduled course of therapy concludes, or, if it suffers from a hardware or software malfunction, it may be taken away from the user. The robot’s absence may, in cases of attachment, cause user distress and possibly result in a loss of therapeutic benefits. Attachment issues can happen with users of all ages, from children to adults and to the elderly. Such issues can be particularly acute in users who cannot understand the causes for the robot’s removal but can arise even with users who have full understanding of the circumstances. Our experiments with SAR robots interacting with elderly users and users with Alzheimer’s disease, mentioned earlier, demonstrate that such users do engage with robots and miss them when the robots are removed [21].

**Perception and Personification of the Robot**

As discussed earlier, one goal of an effective SAR system is to establish a relationship with the user that leads toward intended therapeutic goals. However, since the user cannot be fully informed about the limitations of the robot, the following issue arises: Is there deception inherent in the personification of a robot by a user or a caregiver? Such personification could be unintentional, arising from the caregiver referring to the robot as him or her, ascribing feelings to the robot, and assigning the robot greater intelligence than it may have. Studies have shown that people quickly form mental models of robots they are presented with, much as they do of people. Those models are often incorrect as they are based on what people know best: other people. The designers of the robot may purposefully manipulate the perceptions of the user toward therapeutic goals or may not intend to do so at all; in any case, if such perceptions are incorrect, the user is deceived.

Deception is a risk created by the use of robots in assistive settings. Some roles of SAR systems are most closely associated with people, such as those of a therapist, companion, teacher, or coach. In those roles, the robot may be constructed to physically resemble and act like a human equivalent. In other scenarios, the robot may fill the role of a pet or toy, with physical form to match. While it may be assumed that the physical form of the robot is deliberately designed to evoke the desired type of relationship with the user, there can be unintended ways in which the robot is perceived and received by the user. Studies of the so-called uncanny valley already demonstrate that the level of humanlike realism of the robot has an unexpected impact on people [10]. Similarly, the size of the robot has an impact on the interaction and perceived role: studies have shown that robots that approach the height and size of the user are received with some trepidation compared to smaller embodiments [11]. The way the robot is dressed and accessorized can also influence how it is perceived; a robot in a lab coat and wearing a stethoscope might be perceived as being medically competent even if it is not.

The issues of physical appearance are in many ways just the tip of an iceberg; communication is also crucial. Whether the robot speaks, and if it does so, with a synthetic or recorded voice, male or female, accented or not, and containing emotion or not are all important parameters.
One goal of an effective SAR system is to establish a relationship with the user that leads toward intended therapeutic goals.
Ethical ramifications of SAR are not limited to the balance between risks and benefits. SAR also poses challenges for the user’s informed decision-making ability, as discussed in the next section.

**Autonomy**

The core medical ethics principle of autonomy dictates that patients should be able to make informed decisions about their own care. Extending this principle to SAR, patients should be able to make informed decisions about SAR that are part of their care. As discussed in the previous section, several factors make it likely that a user may not be capable of being fully informed about the abilities and limitations of a particular SAR technology and be aware of his or her own possibly biased perceptions of it. People might believe (or be made to believe) that the robot is more capable than it is, which can create barriers to making an informed decision about care. There are also valid concerns about a user’s privacy with SAR as with most other technologies. If a robot is not able to properly distinguish between confidential information (e.g., personal health information) and information that the user permits for release, then the robot may create an unintended violation of a user’s privacy. In this section, we examine the problems relating to informed consent and privacy that have ethical implications. Since autonomy can also refer to robots that are in control of their own actions, we refer to patient/user autonomy as autonomy while referring to the self-control of a robot as robot autonomy or autonomous robots.

To provide the user with enough information to make an informed decision about a robot, a critical question is: Are the capabilities of an assistive robot being correctly described? If a description of how the robot will be used does not give the user the necessary information to make an informed decision about using the robot, then the caregiver is not behaving in an ethical manner. Consider the example of a companion robot for use in a nursing home that does not allow pets. If the user is told that the robot is just like a pet, but later discovers that in fact the robot only has a limited and small repertoire of behaviors, the user may become disappointed and feel lonely. However, this is not a simple issue; the robot vacuum cleaner, Roomba, is capable of very few actions related to floor vacuuming, yet studies have shown that the users of Roomba are attached to it and demand that it be fixed and returned when broken rather than that it be replaced with a new one [20]. Different users have different expectations, and so it is not necessarily possible to warn a user completely about his/her perceptions and bonding with the robot, positive or otherwise.

Similarly, the role of the robot and possible misconceptions about that role, described in the previous section, could lead a user to expect high-level humanlike medical care from a robot. While the capabilities of the robot may be effective in a specific application domain, they are not comparable to a human doctor or nurse, who may be able to assist the user with decisions or consultations outside of the prescribed therapy. If a user is anticipating an inappropriate benefit for the cost of a robot that she/he is considering purchasing, then that user is not fully informed. The impact of the decision is even more important if the user is considering an application that uses a robot in place of, rather than in addition to, a human caregiver.

The authority of the robot is another sensitive issue for SAR. A robot’s intended role as a therapist may exert influence on the user, putting in question who is in control of the situation and interaction. The question, “Who is in charge?” must be addressed carefully, because the technology may require a level of authority to be effective. A user that is feeling stressed or is in pain must feel free to stop an exercise, for example, even if that is counter to the robot’s advice. However, a SAR system’s role in many contexts is to give direction to a user, requiring some measure of authority derived from expertise. A lack of balance between user autonomy and robot authority could create an ethical dilemma.

When discussing authority with respect to SAR, privacy is of utmost importance. A robot might not have sufficient capabilities to distinguish between privileged information and information that can be distributed. A robot may also lack the ability to distinguish between individuals who have the authority to receive information about the user and those who do not. Patients seeking medical care have an expectation of privacy backed by legal protection. However, a robot might not be able to meet these privacy obligations. In particular, a user might not realize that a robot’s camera could record video, display video in another location, or that wireless transmission of video data cannot be guaranteed to be completely private. People perceive a robot’s camera as having similar capabilities to human vision; this is a natural but false assumption. As discussed in the previous section, the robot might not know to communicate information that is critical to care or how to communicate privileged information in a discreet manner. Therefore, it is important to make sure that the capabilities of a robot are sufficiently explained so that a user has been well informed of a model of the robot’s abilities as possible.

The use of SAR can also have a positive effect on the user’s autonomy. An example from an assistive technology study describes how elders in independent-living situations were asked to allow cameras into their homes to allow for home monitoring for safety. The elders were uncomfortable with this process, as they did not want to be seen, especially in private places like the bathroom. The
Discussion
Preserving the autonomy of a person seeking care is a core ethical value. For the most part, the procedures for informed consent are sufficient for allowing a user’s autonomy in decision making regarding care. However, the potential for user deception can interfere with a user’s informed consent. Currently, the appearance of a robot and its ability to sense its environment and communicate with others might not match. This mismatch might result in (unintentional) deception of the user as to the robot’s capabilities, which in turn may affect the user’s ability to give informed consent. To mitigate this, the users should be presented with a clear description of the robot’s capabilities as well as limitations, but they must also understand that their perceptions of the robot, responses to it, and the attachments and relationships they form with it are not fully predictable, just as they are not in human–human interactions.

Justice
The principle of justice governs the fair distribution of scarce resources. This can be a very difficult topic when discussing experimental treatments such as SAR. The authors know of no SAR systems that are currently used outside the research setting, so discussion of the actual cases in the field is premature. However, we can presume that for the foreseeable future, robots will be somewhat expensive. Thus, a question that should be asked is: Do the benefits of SAR outweigh the costs? Like other proposed therapies, quality of life surveys or other methods for assessing medical economy can be used to assess relative benefits, and costs can be weighed against improvements observed [1], [24]. There does not seem to be a significant difference between calculating the costs and benefits of robots compared to other assistive devices.

Another justice-related issue when discussing robotics in socially assistive settings is the notion of responsibility: Who is responsible when things go wrong? While this might not traditionally pertain to the principle of justice, fair allocation of responsibility for SAR systems might be related to a fair allocation of therapeutic resources. When a robot does not behave as intended, it could be the result of user error or it could be the result of robot error. The difference is not always readily discernable. In the case of robot error, the problem could be in the design, hardware, or software of the robot, meaning that the responsibility belongs to the designer, programmer, manufacturer, distributor, or retailer. Furthermore, the user error may be due not just to a user’s self-imposed mistake but could be a result of poor training, erroneous instructions, or false expectations due to intentional deception.

Software responsibility is troubling since most software licenses explicitly absolve the software developer of responsibility. A large percentage of open-source public domain software and end-user license agreements (EULAs) specify that the software is provided as is and with no liability assumed by the developers or software companies. This includes loss of privacy or data. As privacy is a critical component of the autonomy and nonmaleficence aspects of medical ethics, such a declaration of nonresponsibility is especially concerning. It is entirely possible that a software error could leak privileged information in some way and that the software developer would feel completely within his/her rights to abdicate responsibility for such an error. From the developer’s perspective, software is take-it-or-leave-it. Additionally, a developer cannot be responsible for unforeseen consequences of every line of code, especially given that hardware updates, user error, interface and power issues, and other influences can trigger software errors. This makes the notion of responsibility extremely difficult, making the enforcement of justice related to SAR a challenging prospect, considering that software is just one of the aspects of a complete SAR system.

Discussion
Challenges to the core ethical principle of justice may be the most difficult to anticipate. In fact, most of the problems associated with SAR will be discovered as the robots are used in their target domains. Currently, when robots are tested in research settings with human participants, their use, distribution, and responsibility for errors are all determined by institutional standards, and in the case of many nations, institutional review boards (IRBs). These institutions demand that the inclusion/exclusion criteria, operation of the robot, and responsibility for the robot’s actions be stated in advance. Breaches of such agreements must be addressed on an individual basis, with the termination of a study as a possible consequence. However, as robots are deployed in the consumer realm, similar agreements might not be pursued.

The determination of responsibility for a SAR’s actions is a complex problem that must be addressed, as the technology is being developed and deployed. It is unreasonable to assume that robots will work perfectly or be used always in a completely just and honest manner. Thus,
when breakdowns occur, responsibility and restitution for any harm to a user must be assessed.

Summary
In this article, we have taken the core ethical principles from medicine as a foundation for discussing ethical issues implied by the SAR technologies being developed. Since this ethical framework was constructed with the ethical policies from the United States in mind, and the examples in this article are from North America and Japan, it is possible that different or additional ethical challenges arise for other cultures. More exploration is needed, especially to determine whether robots designed and tested in one medical care system would behave ethically in another. Additionally, as users’ reactions to robots might be different from one group to the next, proven ethical principles for one user population might not be effective for another.

New technologies bring about entirely unprecedented contexts for human–machine interaction and call for thoughtful and well-informed multidisciplinary studies that include inputs and expertise and address concerns from the entire complex constituency, including the technology developers, social scientists, ethicists, and, most importantly, members of the broad user community. This process must be open and ongoing since the technologies and user responses and experiences will continue to evolve indefinitely.

References

David Feil-Seifer, Interaction Laboratory, Center for Robotics and Embedded Systems, Department of Computer Science, University of Southern California, Los Angeles, CA, 90089-0781, USA. E-mail: dfeifer@usc.edu.

Maja J. Matarić, USC Center for Robotics and Embedded Systems, Robotics Research Lab, and Viterbi School of Engineering, University of Southern California, Los Angeles, CA.
It is becoming easier to make robots that seem to understand us and even appear to be like us. There are humanoid robots that can walk, talk, and even shake your hand. There are robots that can recognize human emotional expressions and display emotional signals. There are robots that can recognize particular individuals. There are robot pets that respond to affection and that seem to need looking after. Considerable efforts are being directed toward the development of robots that people enjoy interacting with and want to spend time with. At the same time, developments in robotics are reaching the point where robot caregivers and companions for vulnerable members of society are becoming a real possibility [1], [2]. Before progressing too far down the road toward robot care, it is important to consider what ethical problems are involved in allowing, or even encouraging, the youngest and the eldest members of the society to think that they can form relationships with robots.

The idea of developing robot companions and caregivers for the elderly is taking hold. Elderly people are often lonely and in need of companionship and social contact. Some hold that a robot could be a friend substitute and, at the same time, reassure absent families about the well-being of their elderly relative by monitoring and reporting on their health. Alzheimer’s disease leaves many elderly confused so that they need help with routine activities and someone to answer their questions. It has been suggested that a robot could fulfill this role. Young children need constant care and supervision, but busy
parents do not always have the time to provide it. What harm would there be in a robot nanny taking over some of the care?

Robot companions for the very old and robot nannies for the very young are likely to be designed so that their appearance, movements, and interactions foster the attribution of mental states to them. The aim would be to provide the robots with sufficient features to encourage the target groups to form a relationship with them. Is this a form of deception and is it ethically acceptable? Our focus here is on the ethical issues involved in creating and promoting the illusion of animacy for care robots. We ask what are the pros and cons of encouraging anthropomorphic beliefs in either the elderly or the very young?

**Robot Caregivers and Robot Pets**

The likely development of robot caregivers for the elderly is illustrated by Gecko Systems International Corp.’s predictions that its sales of eldercare personal robots will reach US$8.3 billion by 2014. They are developing the CareBot, a personal robot equipped with multiple vital sign sensors that can follow an elderly person in their home: home-evaluation trials with the elderly began in November 2009. They suggest that the CareBot could become “a new kind of companion that always stays close to them enabling friends and family to care from afar.” The CareBot is capable of verbal interaction, the delivery of medicine, video monitoring, and two-way interactions. Robotsoft’s Kompai robot is similarly proposed as an aide for the elderly. It too can speak and respond to voice commands. Currently, it has an unchanging face, but there are plans to give it the ability to make emotional expressions [3].

Robot caregivers could also be used to look after children, as the Gecko Web site for the CareBot suggests. Other robots have also been developed with childcare in mind. The childcare version of PaPeRo enables mobile monitoring of children. Cameras in the robot’s eyes can transmit images of the child to a window on the parent-care giver’s computer or to their mobile phone. The care giver can see and control the robot to find the child if she moves out of sight. The Hello Kitty robot has a moving head and arms. Despite its limited mobility, the Hello Kitty robot was marketed on some sites as a robot child care giver: “This is a perfect robot for whoever don’t have a lot time [sic] to stay with their child” [2]. [A more recent version (14 September 2010) says, “This is a perfect robot for times when your child needs a little extra comfort and friendship. This Hello Kitty robot will keep your child happily occupied” (http://www.dreamkitty.com).

In addition to robots that are developed with the aim of supervision and monitoring, there has been a considerable interest in the development of robot pets to act as companions. These include Paro, a fur-covered robotic seal, which was specifically designed for therapeutic uses with the elderly. Developed by National Institute of Advanced Industrial Science and Technology (AIST), it responds to petting by moving its tail and opening and closing its eyes. It reacts to sounds and can learn to respond to its name. It makes seallike sounds and is active in the day, preferring to sleep at night. It can detect light and dark by means of a light sensor and recognize when it is being held, stroked, or hit by means of posture and tactile sensors. Sony’s artificial intelligence robotic (AIBO) dog, developed as an entertainment robot, has also been used in robot companions research. It has a metallic doglike form and can walk or chase a ball. It has sensors that can detect distance, acceleration, sound, vibration, and pressure. It can express six emotions (happiness, anger, fear, sadness, surprise, and dislike) by means of its tail, body movements, and the color and shape of its eyes. More recent versions can recognize voice commands, and the robot slightly exhibits different behavior depending on the interactions it has experienced.

Other artifacts have been touted as possible companions for the elderly [4]. Toy robots that could entertain the elderly (or children) include: Pleo, Ifbot, and Primo Puel. Pleo is a robotic dinosaur with many sensors that respond with different behaviors depending on its treatment. Ifbot was developed by Business Design Laboratory Co. for elderly people and can converse with them by means of a large number of stored interaction patterns. Primo Puel is an interactive doll that talks, giggles, and asks for cuddles. It was originally designed to stand in for a boyfriend for young single women but proved unexpectedly popular with elderly women in Japan.

**Anthropomorphism and Deception**

Although some of the robots described earlier have practical purposes, e.g., medical monitoring, most of them have features that persuade people to interact with them and form seeming relationships with them: in other words, to encourage them to be anthropomorphic or zoomorphic toward them.

Anthropomorphism is the term used to describe the behavior of attributing humanlike properties and mental states to nonhuman agents and objects. Zoomorphism is a related concept applied to the attribution of animal characteristics to nonanimals. Robots that move in a human or animal-like way (and/or those that have humanoid or animallike appearances) can encourage anthropomorphism or zoomorphism.

There is a plethora of active ongoing robotics projects aimed at increasing the believability of human–robot interaction. One such area of research is the incorporation of touch sensitivity. The PaPeRo robot has touch sensors on its head and body and can tell if it is being patted or hit. The Huggable [5] has a dense sensor network for detecting the affective component of touch in rubbing, petting, tapping, scratching, and other types of interactions that a person normally has with a pet animal. It seems clear that a robot responding contingent to touch by purring or making pleasing gestures will increase its appeal. For example, Tanaka et al. [6] reported that children were more interested in the quest for curiosity (QRIO) robot that inhabited their nursery when they discovered that petting it on the head caused it to giggle.

Spoken language is a key element in human–robot interaction. Many robots have some ability to recognize and
respond to speech. For example, iRobi (by Yujin Robotics of South Korea) responds to 1,000 words of voice commands. The Kompai robot responds to voice commands and speaks. The Gecko robot responds to voice commands. PaPeRo recognizes about 200 words and gets out of conversational difficulties by making jokes or dancing. Current robots do not have a full-blown natural language-processing interface, yet they can often create the illusion of understanding.

Face recognition is another important factor in developing relationships. Some care robots are already able to store and recognize a limited number of faces, allowing them to distinguish between people and call them by name. An even more compelling way to create the illusion of a robot having mental states and intentions is to give it the ability to recognize the emotion conveyed by a person’s facial expression. Research in emotional expression recognition has been proceeding apace: smile-detection algorithms are incorporated in many digital cameras, and a recent article reports the use of machine-learning methods to distinguish between facial expressions, indicating real or posed pain responses [7]. The development of flexible skinlike materials for robot faces also facilitates their ability to make convincing emotional expressions, as in the Albert Einstein head designed by David Hanson, and augmented with recognition software by the Machine Perception Laboratory at University of California, San Diego.

Robots can be programmed to react politely to us, imitate us, and behave acceptably in the presence of humans [8]. It is possible to make people believe that robots can understand them at least some of the time. Advances in language processing, touch, and expression recognition will act to strengthen the illusion of animacy and sentience and could strengthen human–robot relationships and maintain them for longer.

Should we see efforts to develop features that promote the illusion of mental life in robots as forms of deception? In an important sense they must be, since current robots do not have minds or experiences (in this, we ignore the ongoing debates about whether in future there will be sentient artificial intelligence programs or robots). The question then is, should attempts to create an illusion of robot sentience to foster the belief that a robot is something or someone worth forming a relationship would be viewed as both deceptive and unethical?

Some have argued that this is the case. Robert Sparrow, in particular, has suggested, in the context of a discussion of the possibility of robot pet companions, that any resulting benefits for the elderly, are predicated on mistaking, at a conscious or unconscious level, the robot for a real animal. For an individual to benefit significantly from ownership of a robot pet, they must systematically delude themselves regarding the real nature of their relation with the animal. It requires sentimentality of a morally deplorable sort. Indulging in such sentimentality violates a (weak) duty that we have to ourselves to apprehend the world accurately. The design and manufacture of these robots is unethical in so far as it presupposes or encourages [9].

Sparrow [9] and Sparrow and Sparrow [10] argued that any beneficial effects of robot pets or companions are a consequence of deceiving the elderly person into believing that the robot pet is something with which they could have a relationship. Wallach and Allen [11], in a discussion of the ability of robots to detect basic human social gestures and respond with humanlike social cues, suggest that, “from a puritanical perspective, all such techniques are arguably forms of deception” [11, p. 44].

Should we then conclude that all attempts to induce the illusion of sentience in machines are unethical? We suggest not. Although much of the artificial intelligence depends on creating illusions, and in that sense is a form of deception [12], such a conclusion seems too extreme. The issue of deception is not a straightforward one. It is complicated by the possible anthropomorphic contribution of the viewer. For instance, Zizek [13] describes how people can choose to act as though something were real, “I know very well that this is just an inanimate object, but nonetheless I act as if I believe that this is a living being.” People are anthropomorphic about far more than robots—they often behave as though objects such as their computer or their car were alive (particularly, when things are not behaving as expected). Also, views about artifacts like robots may be unclear—they may be seen neither as being sentient nor as objects but as falling between known categories, as discussed by Turkle et al. [14].

Children enjoy make-believe play and let’s pretend games. As Cayton [15, p. 283] points out, “When children play make-believe and let’s pretend games, they absolutely know it is pretend . . . Real play is a conscious activity. Ask a child who is playing with a doll what they are doing and they may tell you matter-of-factly that they are going to the shops or that the doll is sick, but they will also tell you that they are playing.”

A puppet, on the other hand, is outside of the child’s control and less imagination and pretense is required. But a child left alone with a puppet soon realizes the illusion. The difference with a robot is that it can still operate and act when the child is alone with it. This could create physical, social, and relational anthropomorphism that a child might perceive as real and not illusion. Young children may not know enough about technology to understand the differences between living creatures and convincing robots. The same distinction might be difficult for elderly people with Alzheimer’s disease.

In addition, elements of anthropomorphism may be beyond conscious control. People might report knowing that the robot is interacting with is a machine, but may nonetheless respond to it in some ways as if it were alive. Epley et al. [16] suggest that even metaphorical invocations of anthropomorphism may have an effect on behavior: “Metaphors that might represent a very weak form of anthropomorphism can still have a powerful impact on behavior, with people behaving toward agents in ways that are consistent with these metaphors.”
It may be that a robot that seems to resemble us, or to respond to us, will inevitably be anthropomorphized to some degree. Designing robots to encourage anthropomorphic attributions could therefore be viewed as an unethical form of deception. However, in that case, giving any object a human or animal-like appearance could also be seen as deception. It seems too extreme to suggest that dolls, puppets, and statues should no longer be made or played with. People, in general, and children, in particular, exhibit anthropomorphic behavior much of the time. Anthropomorphic design occurs in many more areas than robotics, from Alessi bottle openers to car grilles and even pet rocks [17]. Rather than objecting to all such uses, it makes more sense to focus our ethical concern on those situations in which anthropomorphic design seems likely to lead to negative consequences for human welfare. Some such consequences are considered in the following section.

A further cause for concern is that there are reasons to expect the vulnerable youngest and eldest members of society to be more likely to be affected by anthropomorphism. Both have a strong need for social contact, and both may lack knowledge of the technology underlying the apparent responsiveness of interactive robots. Both these factors have been argued to increase the tendency to be anthropomorphized in recent accounts [16].

Epley et al. [16] argue that the tendency to anthropomorphize nonhuman agents depends on three psychological determinants: the accessibility and applicability of anthropocentric knowledge, the motivation to explain and understand the behavior of other agents, and the desire for social contact. Their argument is backed up by extensive experimental evidence, of which a few examples are cited here. Various factors can be shown to affect the accessibility and applicability of anthropocentric knowledge: for example, greater similarity between the appearance and behavior of an entity and humans, or animals, can increase the degree of anthropomorphism and empathy shown toward it. Thus, DiSalvo et al. [18] found that robots are anthropomorphized more readily when given human-like faces and bodies. The idea that anthropomorphism is stronger when there is a need to explain is supported by the evidence that shows unpredictability increases the tendency for anthropomorphic explanations [19]. Finally, in accord with the desire for social contact determinant, experimental manipulations show that, when feelings of loneliness are induced, people are more likely to anthropomorphize pets and gadgets [20].

This account of anthropomorphism can be used to argue that both the very young and the very old may be more likely than other age groups to be anthropomorphistic and less able to understand the limited ability of robots to understand and empathize. Both groups have a strong desire for social contact: babies (because they are innately predisposed to look for human social contact) and the elderly (because they are often lonely). In addition, both are likely to lack knowledge about how robots work. Infants and young children are not clear about the differences between living and nonliving entities [21]. Elderly people with Alzheimer’s may not be able to understand the mechanisms underlying robot behavior. Both groups might be more prepared to form relationships with robots and robot pets designed to give the illusion of sentience than other groups of the population.

Both the young and old may show a stronger tendency to anthropomorphize robot companions and pets, but whether or not this is amounts to an ethical problem depends in part on what the consequences of such anthropomorphism might be. We consider these in the following sections.

**Likely Consequences for Robots and the Elderly**

One negative consequence of an elderly person imagining that they have a relationship with a robot might be an increase in their level of anxiety—they might think that they had to look after the robot, even at the expense of their own well-being. Observers and relatives of a confused old person looking after a robot pet might see it as depriving their relative of dignity and infantilizing them.

Similar points have been made in the context of the doll therapy that has been undertaken with those with Alzheimer’s disease. Positive effects have been found from doll therapy, where dolls are given to clients to stimulate memories of a rewarding life role, especially that of a parent, and to act as a focus for reminiscence and conversation [15]. However, ethical objections have been raised to the effect that doll therapy infantilizes the elderly [15].

Studies have shown that clients with dementia engaged in doll therapy tend to believe that their dolls are real babies. When Mackenzie et al. [22] questioned the care workers in homes where doll therapy had been tried, they discovered that some residents would put the doll’s interests before their own as one would with a real baby. They also found that some caregivers, visiting relatives, and fellow residents saw the doll therapy as demeaning and patronizing.

Looking after robot pets could be seen to similarly infantilize elderly people, although a mitigating factor is that robots can be seen as cool gadgets in a way that dolls are not. Another possible negative consequence is that the presence of a robot might result in a reduction in the level of social interaction an elderly person experiences. An outcome in which an elderly person chose to spend time with the robot rather than taking part in social interactions with humans would be unwelcome. Similarly, if other people were to assume that the social needs of an elderly person were being taken care of by the robot and so interacted less with them, that would also be a problem.

On the other hand, there are reasons to expect some positive outcomes. Various studies have found evidence that the elderly can benefit from interacting with robot companions. The positive effects are said to be similar to those obtained from animal-assisted therapy [23]. For instance, Kanamori et al. [24] showed various improvements
in elderly persons who interacted regularly with a Sony AIBO robotic dog—their loneliness scores were reduced, and their quality of life assessment scores improved. Banks et al. [25] even found that elderly people in long-term care facilities benefited as much from interacting with an AIBO robotic dog as from interacting with a real dog. Elderly dementia patients have also shown positive outcomes, including increased communication as a result of sessions with an AIBO [26].

It is of course important to be cautious about the interpretation of such studies. The positive effects depend on comparisons with a control measure. The results reported by Kanamori et al. [24] showed improvements in well-being over time between initial and later sessions. Banks et al. [25] showed that beneficial effects were obtained for those interacting with either the real or the robotic dog, when compared with the control group who received no such opportunities for interaction. However, such improvements could have been found because the alternative was so dire. Someone in solitary confinement might benefit from being given a robot companion, but they would benefit far more from a friendly social environment. It is not clear that the same relative improvements would be found if the comparison were to a control group that received other forms of intervention, such as a visit by someone who chatted and held their hand. It is also important to check that any benefits are maintained over time. An initially interesting robot may rapidly lose its appeal.

Nonetheless, the elderly might obtain some health benefits from interacting with a robot. The robots could also stimulate further social interaction with other people. Robot pets can act as social facilitators, leading to increased interactions between their elderly owners and other people. Robot toys can give an elderly person something to talk about and other people something to talk to them about. For instance, when Wada and Shibata [27] videoed interactions between a Paro robot seal and a group of elderly care home residents, they found that the social interactions between the residents themselves increased at the same time that physiological indicators showed reduced stress levels. It seems that Paro even encouraged positive communication and resulted in a reduction of the backbiting that had previously characterized their interactions.

A robot that facilitates conversation may function as an attractor for visitors. Children may want to play with the robot and have fun with granddad’s big toy. Kanamori et al [24] report the case of an 84-year-old man who talked much more to his children after the introduction of an AIBO robot dog. It gave both him and them a focused object to talk about. In such cases, the underlying deceptive illusion could be justified. Nonetheless, a more utopian vision in which the frail elderly experienced real caring relationships with humans would still seem preferable to a world in which the meaning of their lives depended on animated machines.

Consequences for Babies and Children
Some positive outcomes could result from the combination of elderly people and interactive robots. Positive consequences seem less likely in the case of babies and young infants. Because these youngest members of society have a strong social drive and a lack of knowledge about technology, they are particularly likely to overestimate the abilities of robots that have some of the features of humans or animals. There is a risk that such overestimation by the infants themselves, and by those around them, might result in them spending too much time with robots. This could diminish the time they get to spend in the company of a sensitive human caregiver and impede the development of their understanding of how to interact with fellow human beings.

Infants need to form attachments to a significant caregiver. The kind of attachment they form has a strong influence on their subsequent development. It is well known that, for an infant to become well adjusted and socially attuned, they need a caregiver with sufficient maternal sensitivity to perceive and understand their cues and respond to them promptly and appropriately [28]. It is this that promotes the development of secure attachment in infants and allows them to explore their environment and develop socially. There are disturbing illustrations of the effects of being raised in the absence of human attachment figures in reports of the development of those raised in the impoverished conditions of Romanian orphanges. Nelson et al. [29] compared the cognitive development of young children reared in Romanian institutions to that of those moved to foster care with families. The results showed that children reared in institutions manifested greatly diminished intellectual performance (borderline mental retardation) compared with children reared in families. Chugani et al. [30] found that Romanian orphans who had virtually experienced no mothering differed from children of comparable ages in their brain development and had less active orbitofrontal cortex, hippocampus, amygdale, and temporal areas.

There is little reason to suppose that a robot could provide an adequate replacement for human care. As discussed by Sharkey and Sharkey [2], it is unlikely that a robot would be able to respond to a child in the sensitive manner needed to engender secure attachment. Secure attachment to a caregiver is associated with better development in part because of what the infant learns as a result.

A securely attached child learns to take another’s perspective. When the mother reflects their baby’s emotional distress in their facial expression, it helps the baby form a representation of their own emotions. This social biofeedback leads to the development of a second-order symbolic representation of the infant’s own emotional state [31], [32] and facilitates the development of the ability to empathize and understand the emotions and intentions of others. These are not skills that any near-future robot is likely to have. What patterns of social behavior and reciprocal interaction would a baby learn from a robot that responded contingently to it?

Spending too much time in the company of a robot is unlikely to help and could interfere with an infant’s learning
about the give and take of human relationships. Similarly, a robot is not going to be an adequate replacement for a parent in terms of an infant’s linguistic development. Advances in natural language processing could lead to superficially convincing conversations between robots and children in the near future. However, such interactions would not be meaningful in the way that caring adult–child interactions are. It is one thing for a machine to give a convincing conversational response to a remark or question and a completely different thing to provide appropriate guidance or well-founded answers to puzzling cultural questions. There are many cues that an adult human uses to understand what answer the child requires and at what level.

Language interactions between young children and adults are transactional in nature, both participants change over time. Adults change register according to the child’s abilities and understanding. They continuously assess the child’s comprehension abilities through both language and nonverbal cues and push along the child’s understanding. This is required for both language development and cognitive development in general. It would be extremely difficult to find specifiable rules that a robot could apply for transactional communication to adequately replace a care giver’s intuitions about appropriate guidance.

Babies and infants would probably not be able to resist interacting with a robot that responded to them contingently. There are, however, reasons to fear the effects of such interactions, given that an infant’s experiences of interactions have such a powerful effect on their development. What an infant learns about getting a response from a robot nanny is unlikely to help it understand the subtle and nuanced reciprocal interactions that are needed to form good human relationships. It might seem convenient to have a robot entertain your baby so that you can get some more work done, but the risks might be too great.

In addition to impeded social, emotional, and linguistic development, a young child spending too much time with a robot might suffer other negative consequences. Bryson [33] points out that interactions with robots will be much more predictable than interactions with humans and that children might come to prefer this. In a related argument, Kubinyi et al. [34] argue that just as cross-fostered animals and birds learn behaviors and responses when raised by those of a different species, so humans raised by robots might develop differently. They might, for instance, grow up dependent on individualized entertaining systems and be socialized to follow nonhuman behavioral patterns. A new form of human, homo technicus might emerge [34]. Melson [35] also considers the effects of adapting to pseudo-interactions with technology and suggests that if children begin to think about robots as being alive, they may also begin to think about humans and animals in more mechanistic terms and with less regard to their moral standing.

There are considerable risks of negative consequences from leaving babies and infants in the company of robots. The same is not necessarily true for older children. For children who have formed secure attachments to human caregivers and who have a good grounding in human–social interaction, some exposure to robots might even be useful. Since robots will probably play an increasingly important role in society, it would be just as well if children were educated about their workings and familiar with them. Melson [35] suggests that robotic literacy should be encouraged for both parents and children. “Such ‘literacy’ would help adopters of this technology understand: 1) how robots are produced, maintained, and operated, emphasizing their human-produced properties; 2) what the limits and potentials are for various robotic technologies; and 3) what the distinctions are between living and ‘pretend’ living—stuffed animals, puppets, and robots.” In addition, encouraging children and adults to understand the nature of anthropomorphism and the methods that can be used to strengthen the illusion of mental states in nonliving machines could be a powerful way of protecting them from the ill effects that might result from overestimating the abilities of robots.

**Conclusions**

Clearly, there is a growing interest in developing robot caregivers and companions, particularly for the youngest and eldest members of society. At the same time, there is an ever-increasing ability to implement design features that create the illusion that robots are sentient and able to respond emotionally to us. Such developments raise the likelihood that vulnerable members of society will be left in the company of robots and that people will mistakenly believe that the robots are capable of caring for them and forming mutual relationships. In this article, we have probed the ethics of designing robots that promote the illusion of being able to form meaningful relationships with humans.

It is acknowledged that some form of deception is involved in developing robots that appear to understand us. However, this deception depends on exploiting the natural anthropomorphism of the observer. Anthropomorphic design is prevalent in many aspects of society outside of robotics, and to an extent, being anthropomorphic may be an unavoidable part of being human. Clearly, it would be unreasonable to call all such design unethical. Our arguments are focused on cases where the probable consequences are a reduction in well-being.

It is suggested that, for various reasons, the young and the elderly are likely to be particularly susceptible to such designs. We conclude that robot companions for the elderly could offer positive benefits in terms of improvements in health and welfare, although these are risks in terms of dignity and loss of social contact. In contrast, the development of robot companions and caregivers for babies and infants are more likely to lead to negative consequences. The attachments that infants form with human caregivers fashion the basis of their emotional and social development, and infants that spent too long interacting with robots could learn aberrant forms of interaction. There are reasons to be ethically concerned about the possible effects of exposing either of these vulnerable groups of society to robot care and companionship,
but in the case of the very young, the dangers seem to clearly outweigh any advantages.

References


Amanda Sharkey, Department of Computer Science, University of Sheffield, Sheffield, United Kingdom. amanda@dcs.shef.ac.uk.

Noel Sharkey, Department of Computer Science, University of Sheffield, Sheffield, United Kingdom. E-mail: N.Sharkey@dcs.shef.ac.uk.
This decade has undergone a true robotic demographic explosion. The number of industrial robots in operation exceeded 1 million by the end of 2008. Sales of robots for personal and domestic purposes have increased significantly since 2000 and reached 7.2 million by the end of 2009 [41]. The rampant growth of service robots led to rethink about the role of robots within the human society. Robots are no longer slave machines that respond purely to human requests. They are warranted for some degree of autonomy and decision making. Some, even, envision-friendly and entertaining robots that may become our companions. As a result of this recent robot emancipation, a number of ethical issues have emerged that were not relevant before. We believe that a lively and engaged discussion of ethical issues in robotics by roboticists and others is essential for creating a better and more just world.

In this article, we highlight the possible benefits, as well potential threats, related to the widespread use of robots. We follow the view that a robot cannot be analyzed on its own without taking into consideration the complex sociotechnical nexus of today’s societies and that high-tech devices, such as robots, may influence how societies develop in ways that could not be foreseen during the design of the robots. In our survey, we limit ourselves to presenting the ethical issues delineated by other authors and relay their lines of reasoning for raising the public’s concerns. We show that disagreements on what is ethical or not in robotics stem often from different beliefs on human nature and different expectations on what technology may achieve in the future. We do not offer a personal stance to these issues, so as to allow the reader to form his/her opinion.

In terms of robotic applications, we focus on service robots that peacefully interact with humans [Figure 1(a) and (b)] and lethal robots created to fight on battlefields [Figure 1(c) and (d)]. Other robotic applications are also discussed in the literature; therefore, various concerns for our societies are not discussed here. Unfortunately, for space constraints, we had to limit ourselves in our presentation. For instance, we omitted the question of unemployment caused by the development of industrial robots. This concern is in line with the general issue of using machines to replace human labor, a topic that is central to philosophical debates since the industrial revolution. Furthermore, we chose not to discuss the concerns that robots may one day be able to claim some social, cultural, ethical, or legal rights, that
robots may become sentient machines [51], which we would no longer be allowed to enslave [75], or that we may create robots capable of annihilating mankind [17]. For a discussion on these issues, we refer the reader to [56], [75], and [17].

**Who or What Is Responsible When Robots Harm?**

Veruggio [100], [102] dates the beginning of “roboethics” from two events. One was the Fukuoka World Robot Declaration, wherein it was stated that “next generation robots will contribute to the realization of a safe and peaceful society.” The other was the roboethics road map [101], which sought to promote a cross-cultural discussion among scientists to monitor the effects of robotics technologies currently in use. More recently, an initial sketch of the code of ethics for the robotic community has been proposed [43]. This code offers general guidelines for ethical behavior. For example, the code reminds engineers that they may be held responsible for the actions of artificial creatures that they have helped to design. Along similar lines, Murphy and Woods [70] propose to rephrase the famous Asimov’s laws, which they view as robot centric, in such a way as to remind robotics researchers and developers of their professional responsibilities. For example, the first law was replaced with “A human may not deploy a robot without the human—robot work system meeting the highest legal and professional standards of safety and ethics” [73, p. 19].

All the above implicates the responsibility ascription problem [69]: the problem of assigning responsibility to the manufacturer, designer, owner, or user of the machine when use of this machine led to an armful event is a yet largely open issue. (b) People tend to blame the robots because they falsely attribute them with moral agency [29]. (c) People blame the machine even if they recognize the machine’s lack of free will and lack of intentionality [28]. (d) Many ethicists argue that we should to some extent hold the engineers (the creators of the malfunctioning robots) responsible [60]. (e) To do so, we should use existing the legal principles, or create new ones, if necessary [13].
program, it becomes impossible for the programmer to exhaustively test the behaviors of his/her creations. In other words, the programmer can no longer foresee all possible sets of actions that the robot may take when in function. Hence, the programmer cannot be held responsible if harm should be done as a secondary effect of the robot interacting with humans, as long as the robot was not explicitly programmed to harm people. Matthias suggests that we should broadly adopt the idea of contracting insurances against harm caused by robots. Such a new type of insurance would ensure that, when no one can be held solely responsible for the harm done, then all the people involved in the incident would share the costs.

Marino and Tamburini [60] believe that Matthias's claims go too far. In their opinion, determining who is controlling the robot cannot be a criterion (albeit even the unique criterion) to ascribe responsibility. They argue that engineers cannot be freed from all responsibility on the sole ground that they do not have a complete control over the causal chains implied by the actions of their robots [60]. They rather offer to use legal principles that are routinely applied for other purposes, so as to fill the responsibility gap that Matthias emphasized. They take the example of the legislation in place for ascribing responsibility to the legally responsible person when harm is done by the dependent person. As a result, parents can be held responsible for the act of their children, when they can be found to have not provided adequate care or surveillance, even though there is no clear causal chain connecting them to the damaging events [63, p. 49]. A similar solution is proposed by Asaro [13], who draws a parallel between robots and any other completely unremarkable technological artifact(s) (e.g., a toaster or car). He shows that the Anglo-American civil law that rules for damages caused by these artifacts could also apply to damages produced by robots. For instance, if a manufacturer was aware of the danger that robots create, but failed to notify consumers, he may be charged with a failure to warn. And even if the producer did not know about the danger, he could be accused of failure to take proper care, meaning that the manufacturer failed to recognize some easily foreseeable threat brought upon by his/her technology.

On the downside, Asaro points out that, while the civil law can relatively be easily extended to rule for robot use, the criminal law is hardly applicable to the case of criminal actions caused by robots, as criminal actions can only be performed by moral agents. A moral agent is deemed so when it is recognized capable of understanding the moral concepts conveyed by the bylaws ruling our societies. Without a moral agency, the act of wrongdoing is considered an accident and not a crime. Furthermore, only a moral agent can be punished and reformed. This assumes that the moral agent has the ability to develop and correct its concept of morality [13]. In this context, the responsibility-ascription problem is, hence, reduced to the issue of attributing moral agency to the robot. Several authors have approached the problem of ascribing moral agency to robots [91]. For instance, Harnard [37] proposes to use some sort of moral Turing tests to establish whether the robot can be held responsible in court.

Another issue around the responsibility ascription problem centers on attributing moral agency to a robot. In one study, Friedman and Millett [30] found that 83% of the undergraduate computer science majors they interviewed attributed aspects of agency, either decision making or intentions, to computers. In addition, 21% of these students consistently held computers morally responsible for errors. In another article, Friedman and Kahn [28] identified a situation that may increase peoples attribution of agency to a machine, namely, when the machine is an expert recommendation system. Friedman and Kahn provide an example of the acute physiology and chronic health evaluation (APACHE) system [21]: a sophisticated computer-based modeling recommendation system to help hospital staff determine when to end life support for patients in intensive care units. Friedman and Kahn argue that the more such a system is relied on for objective and authoritative information, the more difficult it becomes to override its recommendations, and the more likely staff, including physicians, could begin to attribute moral agency toward the system. As a potential solution to such problems, Friedman and Kahn offer two design strategies. First, computational systems should be designed in ways that do not denigrate the human user to machinelike status. Second, computational systems should be designed in ways that do not impersonate human agency by attempting to mimic intentional states. The problem, however, in applying this second recommendation to robot design and implementation, especially those robots that have a humanoid form, is that such robots by design are conveying human attributes, thus fostering this problem.

Ethical Issues in Service Robots

The design principle mentioned in the previous section aims at ensuring that robotic systems remain easily distinguishable from humans. Accordingly, this principle should help people ascribe responsibility in cases when the machine malfunctions or harms someone. However, as we noted, the current trend in robotics is the opposite, as there is a growing effort to design robots so that they look like humans [44], [45] or animals [31], [89]. The idea of designing machine-masquerading humans was questioned by Miller on the ground of human freedom [67]. Miller argues that, if humanlike robots really came to share the human space on a daily basis, the humans should be allowed to decide whether they wished to interact with these creatures; if they should decide they wanted to
interact solely with the other humans, they should be given the freedom to do so. Similarly, efforts at endowing robots with social skills have been criticized on the ground that the number of meaningful social interactions that humans that are typically capable to maintain is relatively small [23], [47]. Therefore, interacting with social artificial agents on a regular basis may lead people to become less prone to engage in social interactions with other people [66]. Others even hypothesized that people may come to build strong and perhaps even intimate bounds with robots and that this, again, may have negative side effects on the emotional relationships that people may be able to build with other people [50].

To shed some light on the aforementioned debate, people have started studying the type of human–robot relationships that arise when people interact with robotic systems that mimic human or animal behavior. In a series of four studies, Kahn and his colleagues studied children’s social and moral relationships with the robot dog, the artificial intelligence robot (AIBO). The first three studies compared children’s interaction with and reasoning about AIBO to, respectively, a stuffed (nonrobotic) dog [49], a biologically live dog [65], and a mechanical nonrobot dog [94], whereas the fourth study analyzed over postings in AIBO online discussion forums that spoke of members’ relationships with their AIBO [30]. Together, these four studies provide converging evidence that children and adults can and often do establish meaningful and robust social conceptualizations and relationships with a robot that they recognize as a technology. For example, in the online discussion forum study, members affirmed that AIBO was a technology (75%), lifelike (48%), had mental states (60%), and was a social being (59%).

Across these four studies, however, the researchers found inconsistent findings in terms of people’s commitments to AIBO as a moral agent. In an online discussion forum study, e.g., only 12% of the postings affirmed that AIBO had moral standing, including that AIBO had rights, merited respect, engendered moral regard, could be a recipient of care, or could be held morally responsible or blameworthy [30]. In contrast, in the Melson et al.’s [65] study, it was found that while, on the one hand, the children granted greater moral standing to a biologically live dog (86%) than to AIBO (76%), it was still striking that such a large percentage of children (76%) granted moral standing to the robot dog at all. One explanation for these inconsistent findings between studies is that the measures for establishing moral standing have been few and themselves difficult to interpret. For example, two of the five moral questions in the Melson et al.’s study were as follows: If you decided you did not like AIBO anymore is it OK or not OK to throw AIBO in the garbage? and If you decided you did not like AIBO anymore is it OK or not OK to destroy AIBO? The “not OK” answers were interpreted as indicating moral standing. However, one could plausibly make the same judgment about throwing away or destroying an expensive computer (because, e.g., it would wasteful) without committing morally to the artifact [65].

Since humans can develop emotional attachment toward robots, concerns have been expressed regarding the long-term consequences that such attachment may have on the individual. This is especially relevant when the person is fragile, as it is the case with children and people with mental delays. However, there are also several reasons to rather believe that interacting with social robots may benefit some of these individuals [48], [54], [97]. For instance, interacting with robots that display social behavior may help children with autism-impaired social skills [80], [26]. Robins et al. [80] conducted longitudinal studies over the course of several weeks of children with autism interacting with a humanoid robot. Unknown to the children, the robot was puppeteered so that it imitated the children’s movement. Robins et al. showed that repeated exposure to the robot facilitated the emergence of spontaneous, proactive, and playful behavior, which these children very rarely display. Furthermore, once accustomed to the robot, the children also seem to engage in a more proactive interactive behavior with the adult investigator present in the room during the experiment. This leads, in some cases, to a triadic interaction: child–robot–adult. For example, children would acknowledge the presence of the investigator by spontaneously sitting on his/her lap for a few moments, holding his/her hand, or even trying to communicate by using simple words. However, it was not clear whether the social skills that children exhibited during the interactions with the robot had lasting effects.

In another study, Feil-Seifer and Mataric used a bubble-blowing robot in a three-some interaction child–caretaker–robot. While the robot was not actually behaving socially, its automatic bubble-blowing behavior provoked more child-caretaker interactions. In a similar triadic child–parent–robot scenario, Kozima and colleagues conducted a series of studies using Keepon, a simple two-link robot ball face, whose motions conveyed emotional expressions. These studies comfort Robins et al.’s findings that children with autism, in such a triadic scenario, spontaneously engage in social and affect display, which they otherwise tend to avoid [55], [26]. A comparative study of children with autism interacting with AIBO as opposed to a simpler mechanical toy showed enhanced verbal address directed to AIBO [94]. A survey of these studies can be found in [79].

As a whole, these studies seem to indicate that playing with robots that appear to behave in an autonomous and social manner may help children with autism-impaired more of these social skills that the autism therapy seeks to promote. Such a robotic-aided therapy does not aim
at developing attachment of the children toward the robot, but it might be a potential side effect. The question remains whether it is ethically correct to encourage children with autism to engage in affective interactions with machines incapable of emotions. Dautenhahn and Werry's response is that, "from the perspective of a person with autism and his/her needs, are these ethical concerns really relevant?"

Similarly, robotic pets used in therapy with elderly may offer some level of companionship. The seal robot, Paro, is probably the best example of such an application [89] [Figure 1(b)]. Wada et al. [104] reported on an extended use of Paro as part of therapeutic sessions in pediatric wards and elderly institutions worldwide. The results showed that the interaction with Paro improved the patients' and elderly people's moods and reduced their stress level [103]. It made them more active and communicative both among themselves and with their caretakers. A pilot study using electroencephalography (EEG) suggested that this robot therapy may improve the pattern of brain activity in patients suffering from dementia [104]. Furthermore, the effects of long-term interaction between Paro and the elderly were found to last for more than a year [105].

Although the aforementioned results speak in favor of using robots for therapy with the elderly, Sharkey offers a more cautious argumentation [85]. In his opinion, such surrogate companions do not really alleviate the elderly's isolation, and people are deluded about the real nature of their relationship to the devices [92] (Figure 3). Furthermore, even the robots that are clearly helping the elderly to maintain independence in their own homes [27] (e.g., robots used to remind the patient to take his/her medication) could lead to a situation where the elderly is left exclusively to the care of machines. However, the elderly's mental health substantially depends on human contact, which is to a large extent provided by the caregivers [93].

Robot nannies are another example of robotic applications that raise ethical questions [88]. There is an effort, mainly in South Korea and Japan, to build more sophisticated robots that could not only monitor babies [e.g., personal partner robot by National Electronics Conference (NEC) [32], Figure 1(a)] but would also be equipped with enough autonomy so as to call upon human caretakers only in unusual circumstances. It is likely that children will spend time playing with child-care robots, as researchers progress in designing ways for the robot to offer a sustained and rich interaction with the child, which may span months or even years [51], [63], [88]. This may, however, be detrimental to the physical and mental development of the child if children were to be left without human contact for many hours per day, as currently robotic pets are not designed to participate in the child's development in the same way as a child minder is trained to look after children [85].

This remains very speculative as the psychological impact that such robotics care may have on children's development is unknown. Some attempted to draw parallels with reports on severe social dysfunctions in young monkeys those interacted solely with artificial caretakers throughout the first years of development [61], [16], [88]. Perhaps of more pressing concern is the fact that there is no regulation to specifically deal with the case of child abuse when the child is cared for by a robot (national and international laws protecting children from mistreatment such as the United Nations Convention on the Rights of Child [71] do not cover this case) [88]. While one may argue that, when the time will really come to see robots caring for children, one will work on the associated legal issues, some people counter that this may be a bigger challenge than expected, as providing a unified code of ethics for regulating the use of robot nannies may be impossible owing to cultural differences between nations [36].

**Ethical Issues in Lethal Robots**

In the previous section, we discussed some of the ethical issues that stem from the current or foreseen robotic applications of service robots for education and therapy. Of equal if not more immediate ethical concerns are the current military applications of robots. Even though fully autonomous robots are not yet running in battlefields, as we will discuss here, the risks and benefits that introducing such autonomous lethal machine may have on wars are of crucial importance. Furthermore, because military technology often finds its way into civil applications, such as security or policing [14], [87], discussing the ethical issues related to military robots might also serve a broader context.

Currently, the decision to use a robotic device to kill human beings is still taken by a human operator. This decision stems from the desire to make sure that the human remains "in the loop," but it is not made out of technical
necessity [14]. It is clear that the margin that separates us from having fully autonomous-armed systems in the battlefield is thinning. Even if all armed robots were to be supervised by humans, one may still wonder to what extent the human is still in control [9]. Moreover, there may be cases where one cannot avoid giving full autonomy to the system. For instance, combat aircrafts must be fully autonomous to effectively operate [99]. Sharkey predicts that, as the number of robots in operation in the battlefield increases, they may outnumber human soldiers. He then argues that it will become impossible for humans to simultaneously operate all these robots. Robots will then have to be fully autonomous [83].

One ethical issue (perhaps the issue that received most attention to date) arising from increasing autonomy of war robots has to do with the problem of discriminating between the fighters and innocent people. This distinction is at the core of the just war theory [106] and humanitarian laws [82]. These laws stipulate that only the fighters are legitimate targets and prohibit attacks against any other nonlegitimate targets [84], [14]. Sharkey rightfully argues that our robots are yet far from having visual capabilities that may allow to faithfully discriminate between the legitimate and nonlegitimate targets, even in close-contact encounter [85]. Besides, distinguishing between the legitimate and illegitimate targets is not purely technical and is further complicated by the lack of a clear definition of what is a civilian. (The 1944 Geneva Convention advises to use common sense, and the 1977 Protocol I defines a civilian as any person who is not a fighter [72].) However, even if one was provided with a precise definition that could be encoded in a computer program, it is doubtful that robots would achieve, in a foreseeable future, a level of complexity in robot cognition that would allow the robot to recognize ambiguous situations involving a nonlegitimate target manipulating lethal instruments (such as a situation where a child is carrying guns or ammunition). Sharkey argues that autonomous lethal systems should not be used, as long as one cannot fully demonstrate that the systems can faithfully distinguish between a soldier and civilian, and this in all situations [83]. Lin et al. believe that this is too stringent a condition, since even humans make errors of this kind (Figure 4) [58]. Arkin counters that, although unmanned robotic systems may make mistakes, it would on an average behave more ethically than human beings [9]. In support of this, Arkin cites the report from the Surgeon General’s Office [96] regarding the ethics of soldiers. Less than half of the soldiers believed that the nonfighters should be treated with dignity. The other half was unclear as to how they should be treated. Moreover, one tenth of interrogated soldiers had mistreated nonfighters and one third reported having at least once faced a situation where they felt incapable of deciding what was the correct action (although all soldiers had received ethical training). Since human soldiers appear to misbehave from time to time, using machines that are more reliable and hence would, on average, make less mistakes should bring more good than harm. Lin et al. share the view that human soldiers are indeed less reliable and report on an evidence that human soldiers may act irrationally when in fear or stress. Hence, they concur that combat robots, which are affected by neither fear nor stress, may act more ethically than human soldiers irrespective of the circumstances [58].

Lin and colleagues point to one more issue related to using combat robots. As in the case of any other new technology, errors and bugs will inevitably exist, and these will lead combat robots to cause harmful accidents [58]. Such bugs or errors will be far more costly as human lives might be at stake. They advise to perform extensive testing of each military robot before usage. Nevertheless, they anticipate that, regardless of such efforts, combat robots may still occasionally behave in unexpected or unintended ways when used in the actual field [58]. Such errors could even lead to accidental wars if the robot’s unexpected aggressive behavior was to be interpreted by the opponent as an act of war [14]. Groups of people interested in starting a war may seize upon such accidents to justify hostilities.
Even if one is not disputing the ethical question of fighting a war, one may want to question the ethics of having armed robots fully autonomous and used routinely in battlefields, especially when only one side may have robots. Politicians may tend to favor efforts made to replacing human fighters with robots, as each country feels a moral obligation to protect the lives of its soldiers [83]. However, there may be long-term consequences of waging these so-called risk-free wars (“A war where pilotless aircraft can beat a country’s forces before sending in the ground robots to clean up” [87, p. 16]) or push-button wars (“A war in which the enemy is killed at a distance, without any immediate risk to oneself” [15, p. 62]). Since such wars will return wrecked metal instead of dead bodies (at least to the country using only robots), the emotional impact that wars currently have on civilians of that country will be largely lessened. The above is true only for the civilians not affected directly by combat, i.e., for wars fought in a distance.

It is feared that this may make it easier for a country to launch a war. These wars may also last for longer periods of time [58]. There are contradicting opinions whether this may result in people growing indifferent to the conduct of war. Sharkey fears that this would be the case [83], whereas Asaro believes that people are nearly always averse to starting an unjust war, irrespective of whether it would lead to human fatalities [15, p. 58]. That the war is risk free does not make it more acceptable [14]. Lin et al. counterweight this line of reasoning, arguing that such reasoning may lead to even more dangerously foolish ideas, such as the idea of trying to prevent wars to happen by increasing the brutality of fighting [58].

It was also argued that risk-free wars might increase terrorism, as the only possibility to strike back on a country that uses mainly robots in wars is to attack its citizens [83]. The less advanced, technologically speaking, side may advocate terrorism as a morally acceptable means to counterattack on the ground that robot armies are the product of a rich and elaborate economy, and that the members of that economy are the next-best legitimate targets [15, p. 64]. Hence, risk-free wars may paradoxically increase the risks for civilians [46]. However, Asaro reminds us that the wars are deemed morally acceptable as long as they do not harm civilians. According to this definition, terrorism would not be justified, irrespective of whether it is meant as a response to a country using robot armies. Thus, the fear that terrorism may increase as a result of using robot armies does not constitute, in Asaro’s view, a valid moral objection to using robot armies. Only the questions of whether the robot armies can cause more harm or whether the use of such armies may lead to unjustified wars are of essence in the debate [14].

In contrast, Arkin anticipates that we will not end up with armies of unmanned systems operating on their own, but that rather heterogeneous teams composed of autonomous systems and humans soldiers will work together on the battlefield. He expects this to become a standard. Wars would, hence, not be fully risk free and so the dreaded consequences in increased terrorism or in societal indifference are not to be feared. Furthermore, Arkin expects that mixed teams, composed of robots and human soldiers, will act more ethically than groups composed of solely human soldiers. Robots equipped with video cameras (or other sensors) will record and report actions on the battlefield. Thus, they might serve as a deterrent against unethical behavior, as such acts would be registered. However, Lin and colleagues argue that if soldiers were to know that they are being watched by their fellow robot soldiers, they may no longer trust them and this could impact team cohesion. Consequently, human soldiers may fail to act adequately, e.g., by not providing support even if it is justified, out of stress caused by constant monitoring [58].

Lastly, Sharkey points out that the legal status of war robots is unclear [86]. For example, while the unmanned aerial vehicle RQ-1 Predator [Figure 1(d)] was developed as a reconnaissance machine (hence the R in the name), it was subsequently equipped with hellfire missiles and renamed MQ-1 (where M stands for multipurpose). The MQ-1 was, however, never approved as a weapon. The fact of utmost concern is that, under current military standards, the MQ-1 does not need to be approved. Since the bare RQ-1 was not considered as a weapon (since it was meant only for surveillance) and that hellfire missiles have already been approved separately as weapons, the combination does not need special approval [19]. This may create a precedent whereby armed robots with growing level of autonomy can be created and used without any real legal control. In relation to legal issues, Asaro notes that “what is and what is not acceptable in war” is ultimately the subject of convention between nations [15, p. 64]. He argues that we can find support in existing laws only to certain extent. Eventually, the international community will be forced to create new laws and treaties to regulate the use of autonomous fighting robots.

**Machine Ethics**

Although still in its early stages, machine ethics offers a practical approach to introducing ethics in the design of autonomous machines. Machine ethics aims at giving the machine some autonomy, while ensuring that its behavior will abide ethical rules. Primarily, machine ethics seeks methods not only to ensure that the machine’s behavior toward humans is proper [4], but it may also extend to designing rules driving ethical behavior of a machine toward another machine [6]. Machine ethics extends the field of computer ethics that is concerned with how people behave...
with their computers to address the problem of how machines behave in general [2].

The interest in machine ethics is driven by the fact that robots have been already tightly integrated into human societies. Thus, since the robots already interact with humans and, as argued in the section “Who or What Is Responsible When Robots Harm?” engineers could be held responsible (to certain extent) for the actions of their creations; it is desirable to find methods of equipping the machines with moral behavior. Importantly, although the public attention might be focusing on the military application (such as Arkin’s military adviser providing guidance on the use of lethal force by a robot [11]), machine ethics seems to be more concerned with service robots. There are many examples of such applications. Robots that share the workbench with humans in the industry might no longer be considered just a manufacturing tool but also as a “colleague” with whom workers interact [20]. Artificial sales agents in e-commerce, which can predict customers behaviors, should not abuse this knowledge by displaying unethical behavior [39]. Driverless trains in extreme situations might be forced to make decisions that could have life or death implications [2].

Asimov’s laws of robotics are one of the first and best-known proposal to embed ethical concepts in the controller of the robot. (Asimov’s laws of robotics were first introduced in the short science-fiction story Runaround [15].) According to these, all robots should under all circumstances obey three laws:

1) A robot may not injure a human being or, through inaction, allow a human being to be harmed.
2) A robot must obey orders it receives from human beings, except when such orders conflict with the first law.
3) A robot must protect its own existence as long as such protection does not conflict with the first or second law.

Later, Asimov added the fourth law (known as the law zero).

4) No robot may harm humanity or, through inaction, allow humanity to come to harm.

Many researchers recognize that Asimov’s laws assume that robots have sufficient cognition to make moral decisions in all situations, including the complicated ones, in which even humans might have doubts [70]. Consequently, keeping in mind the current level of AI, these laws, although simple and elegant, serve no useful practical purpose [9] and are thus viewed as an unsatisfactory basis for machine ethics [8], [34]. Nevertheless, Asimov’s laws often serve as a reference or starting point in the discussions related to machine ethics.

Fedaghi [1] proposes a classification scheme into ethical categories to simplify the process by which a robot may determine which action is most ethical in delicate situations. As a proof of concept, Fedaghi applies this classification to decompose Asimov’s laws, hereby showing that these laws, once rephrased, can support logical reasoning. Such an approach is in line with the so-called procedural ethics [59], which develops procedures to guide the process by which ethical decisions are made [1]. A similar approach is presented in [18] that draws inspiration in Gottfried Wilhelm Leibniz’s dream of a universal moral calculus [60]. There, deontic logic [22], [68] (i.e., logic extended with special operators for representing ethical concepts) is used instead of Asimov’s laws to ground the robot’s ethical reasoning. Such a methodology aims at maximizing the likelihood that a robot will behave in a certifiably ethical manner. That is, the robot’s actions will be determined so that the ethical correctness of the resulting robot’s behavior can be ensured through formal proofs. Such formal proofs check if a given robot 1) only takes permissible actions and 2) performs all obligatory actions (subject to ties and conflicts) [12]. Promoters of such methodology reason that human relationships and by extension human–robot relationships need to be based on some level of trust [107]. Such a formal and logical approach to describing robot behavior may help in determining whether the system is trustworthy. In contrast, they view inductive reasoning, which is based on case studies, as unreliable, because, while the “premise (success on trials) may all be true, the conclusion (desired behavior in the future) might still be false” [18], [90].

Others oppose this point of view and advocate the use of case-based reasoning (CBR) [74]. They reason that people can behave ethically without learning ethics (drawing a parallel to the fact that one can speak fluently a language without having received any formal grammar lessons) [81]. For example, McLaren implemented a CBR-ethical reasoner [64] and Anderson created a machine-learning system that automatically derives rules (principles) from cases provided by an expert ethicist [3], [7], [5]. For example, Arkin uses deliberative/reactive autonomous robotic architectures and provides the theory and formalisms for ethical control [10] and applies these to automatic military advisor [11]. He considers stimuli to behavior mappings and extends them with ethical constraints to ensure appropriate robot response (consistent with the law). In another example, Honarvar [40] used a CBR-like mechanism to train an artificial neural network to classify what is morally acceptable in a belief–desire–intention framework [77]. For example, he used this framework to augment the ethical knowledge of sales agent in an e-commerce application [39].

A particular machine ethics system that is very easy to implement is the one based on utilitarianism. It uses mathematical calculus to determine the best choice (by computing
and maximizing the goodness, however defined, of all actions) [4]. However, since utilitarianism values benefits brought upon society as a whole, hence ignoring the fate reserved to each individual in the society [78], such moral arithmetic cannot protect the fundamental rights of each individual [11] and as such is mostly of limited interest [35]. Still, practical work with a certain utilitarian flavor can be found in the literature, as most CBR systems previously presented assume that an arithmetic value is the main basis for determining what it is moral to do [53].

The last approach that we will mention is the rule-based one proposed by Powers. Powers argues that ethical systems such as Kant’s categorical imperative naturally lead to a set of rules. (A categorical imperative denotes an absolute, unconditional requirement that asserts its authority in all circumstances, e.g., “act only according to that maxim whereby you can at the same time will that it should become a universal law” [55, p. 30].) This approach, hence, assumes that an ideological ethical code can be translated into a set of core rules. This is slightly similar to the deontic logic we reviewed earlier. It allows the robots to logically derive new ethical rules, appropriate to particular and new situations.

Although interesting, this approach has not gathered much attention, as researchers usually turn to pure logic systems or CBR. In addition, Powers’ ethical system had been criticized by Tonkens [98] on the basis that the development of Kantian artificial agents is itself against Kant’s ethics. According to Kant, moral agents are both rational and free, whereas machines can only be rational. Hence, the mere fact of implementing a sense of morality into machines limits the machine’s freedom of thought and reasoning.

In conclusion, machine ethics is composed of a number of interesting attempts to embed ethical rules in the robot’s controller. These may be either popular ethics rules, such as Asimov’s laws, or derived from classical philosophical approaches to ethics, such as Kant’s ethics. Logical reasoning is the driving framework for most approaches. While still in infancy, machine ethics is a valuable attempt to conciliate the need to provide robots with ethical behavior with the need to make these machines more autonomous, as they come to support humans in their daily life. However, the approach may fall prey to several problems discussed throughout this article. Three of those stand out. One, if machines are not capable of being moral agents, as most philosophers agree, then it is important to design them with the ability to make moral decisions. Second, equipping the machines with morality (assuming it is possible) does not need to be a moral act on its own and might depend on the application one has in mind while developing a moral robot. For example, embedding morality into robot nannies or combat robots could lead to their widespread use, which could have severe negative consequences on the society. Finally, in an attempt to embed ethics into machines, because of their limited cognition, one must often unduly simplify the moral life. This seems to stand against the very goal of machine ethics itself (at least to some extent). It seems that it is still too early to judge whether the methods of machine ethics will prove useful or not and await more applications implemented in life.

Conclusions

Almost everyone agrees that they want robots to contribute to a better and more ethical world. The disagreements arise in how to bring that about. Some people want to embed ethical rules in the robots controller and employ such robots in morally challenging contexts, such as on the battlefield. Others argue vehemently against this approach: that robots themselves are incapable of being moral agents and thus should not be designed to have moral decision-making abilities. Others want to leverage the social aspects of robotics in bringing about human good. Along these lines, researchers have explored how robots can help children with autism or assist the elderly physically, thereby provide the elderly with enough autonomy to allow them to live in their own residence. Other researchers have explored how robots can provide companionship for the elderly and general population. Still others have worried that no matter how sophisticated robots become in their form and function, their technological platform will always distinguish people from them and prevent depth and authenticity of relation from forming. These are all open questions. Some are philosophical in nature, as is the question of whether robots are moral agents or could be in the future. Some are psychological, as in the question of whether people attribute moral responsibility to robots that harm. Some require political answers and new legislation. Finally, some, if not many, of the questions require thoughtful and on-going responses by those who engineer and design the robots. The engineer is also responsible for the ethical consequences of his/her creation. This seems at odds with the way research is currently done in robotics. Rarely, does one question the long-term ethical consequences of the research reported upon in scientific publications. (We are not referring here to short-term ethical consequences of a research, such as a research that involves human subjects. Clearly, these are always carefully scrutinized, and this research must be approved by the ethical committee before the conduct of the project.) There are several reasons for this. On the one hand, most of these damaging long-term consequences seem very speculative and still far away from the technological reality. On the other hand, it is expected that these issues will be disputed at a political level, and, hence, that it is perhaps not the role of the engineers and scientists to discuss these.

Some scientists, however, discuss these issues, but, as with any debate, people sometimes have opposite views on which robotic application is ethical and which is not. We showed that such dissensions stemmed often from different beliefs on human nature and different expectations on what technology may achieve in the future. Although it is difficult to anticipate how and when robots will come to play an active role in our society, there is no reason why one should
not continue discussing various scenarios. We might be motivated by the beauty of our artifacts, their usefulness, or the economic rewards. However, in addition, we are morally accountable for what we design and put out into the world.

Acknowledgments
This work was partially supported by the Swiss National Science Foundation (grant number K-23K0-117914), the European Commission under contract number FP7-248258 (First-MM), and the National Science Foundation in the United States (grant number IIS-0905289).

References


Pawel Lichocki, Laboratory of Intelligent Systems, Ecole Polytechnique Federale de Lausanne, 1015 Lausanne, Switzerland. E-mail: pawel.lichocki@epfl.ch.

Peter H. Kahn, Jr., Department of Psychology, University of Washington, Seattle, WA 98195-1525. E-mail: pkahn@u.washington.edu.

Aude Billard, Learning Algorithms and Systems Laboratory, Ecole Polytechnique Federale de Lausanne, 1015 Lausanne, Switzerland. E-mail: aude.billard@epfl.ch.
Remote-Control Crimes

By Peter M. Asaro

Roboethics and Legal Jurisdictions of Tele-Agency

This article considers some of the potential legal implications of teleoperated robotic systems for enabling action at a distance or tele-agency. In particular, it considers issues that may confront law enforcement as well as issues of legal jurisdiction when tele-agency extends across the traditional physical boundaries of legal jurisdiction.

The legal approach is one of the approaches for the issues faced by roboethics. A consideration of ethics in robotics using the tools offered by the practice of law has been made elsewhere [1]–[3] and has focused on product liability and robots as legal agents. The difficulties of applying the law to some of the possible activities involving new robotic capabilities that may arise in the near future are considered.

One new capability, in particular, is that robotic systems also pose one of the greatest threats of social disruption. This new capability has, however, been largely overlooked by the rather small literature on roboethics, namely, the ability of robotic systems to support action at a distance, known as tele-agency. This capability has serious implications for both law enforcement and legal jurisdiction, though tele-agency has received more attention in the art world (e.g., [4], [5]) to date than it has in discussions of robot ethics and law. This essay seeks to correct this by considering some of the legal issues that might arise as teleoperated robots proliferate and spread into consumer markets, international trade, and hacker communities.

Telecrimes and Law Enforcement Issues

Simply put, teleoperated and remotely controlled systems allow the legally responsible actor(s) in control of the system to be spatially (and, in the case of preprogrammed systems, temporally) distant from the effects of their actions, without requiring the support of human accomplices. This has several serious consequences for law enforcement because of the perpetrator’s reduced bodily risk, the risk of being arrested in the conduct of a crime, and the difficulties
involved in correctly identifying the responsible individuals. These are clearly matters of law enforcement—identifying, apprehending, and convicting the perpetrator of a crime—and do not affect whether or not the perpetrators are guilty of a crime. However, the ability to more easily commit crimes while reducing the risks of facing punishment is certainly a threat to justice and the public good.

While, in theory, there is little legal difference between robbing a bank at gunpoint and using an armed robot to rob a bank, there are significant practical differences. Most obviously, there is a significant difference in the bodily risks assumed by an armed robber and someone controlling an armed robot remotely, which has implications for the use of armed police and security guards as a deterrent to such crimes. Viewed another way, this could have a significant impact on the ability of police and security forces to intervene and stop such crime, especially if they are unable to physically subdue the robotic system. On the other hand, law enforcement could use force against the robot without the same restraint that would be called for if a human body were at risk (even the body of a suspected criminal).

The use of such a remote-robbing robot also requires the police to do additional work to correctly identify and locate the perpetrator of the robbery. If the current state of the art of cybercrimes is a good indication, it will likely be quite difficult to track down the perpetrator of such a crime when their control path has been routed through a series of networks and servers intentionally designed to obscure the identity and location of the criminal. Teleoperation implies the real-time control of a robot from a physical distance. We can, however, also consider a preprogrammed robot as a kind of teleoperation in which the programmer/controller is temporally removed from the actions of the robot, though it is still the responsible agent. Again, there are precedents for treating programs as a form of criminal behavior, as is done in the creation and use of illegal viruses and botnets.

Robotic technologies might not prove to be a source for a massive remote-control crime wave, however. Some reasons for this are that, first, in material-property theft, the stolen property must be recovered at some point, and thus the police could track and follow the robot until the perpetrators attempt to retrieve the property. Telerobotic theft is quite unlike cybertheft in this regard, as stolen information can be quickly and easily transferred through the network, while stolen material objects cannot. Second, initially such robotic technologies will be expensive and thus would be unlikely to be used in petty crimes where the value of the stolen goods is less than the cost of losing the robot. Third, like other cybercrimes, these tele-agency crimes might leave data trails that could be used to identify perpetrators, and videofeeds and control commands might actually be recorded by authorities in ways that could be used in courts to prove the guilt of perpetrators. While it is already possible to commit complex cybercrimes, robotic technologies will extend the range of these crimes into the embodied material world, including bodily violations and violent crimes such as assault, rape, and murder.

**Tele-Agency Across Jurisdictions**

Complex legal questions may also arise when the perpetrator controlling the system and the robot being controlled are in different legal jurisdictions. Certain interjurisdictional or multijurisdictional actions are already handled by the law in various ways. Examples include using other human agents to conduct a crime, such as in conspiracy or being an accomplice to a crime, though these tend to carry lesser penalties than the actual commission of the crime. In the United States, a crime (e.g., a fraud) that involves actions in two or more different states within the country can result in the matter being settled by the federal court system rather than in the state courts. Prosecutors may also seek convictions for crimes in each state jurisdiction separately, depending on the case and cooperation between state and federal prosecutors. In these cases, there are often similar sets of laws that apply in each jurisdiction, and sometimes one set overrides, such as federal law having precedence over state, provincial, or local laws. More controversial are cases in which there are subtle differences in the definitions of what constitutes the crime in question or when different penalties may apply, depending on the court and jurisdiction in which the trial is held. For example, a first-degree murder in some U.S. states carries a death penalty, but not in others, and so it can matter a great deal where the crime is committed and tried. Indeed, the rules of extradition in some jurisdictions, as in many countries in Europe, are such that they will not allow the extradition of a suspect for trial in another country in which they might face the death penalty. More generally, extradition requires an international treaty agreement, which usually stipulates that the charges be serious enough to warrant returning an individual, such that many petty crimes committed using robots might not warrant extradition.

The more problematic cases involve activities that are legal in one place but illegal in another. A good example of this is gambling. In most states, gambling is illegal, or at least tightly regulated by the state. With the advent of the Internet, however, it became possible to engage in gambling activities online. The legal question then arose as to where the gambling is taking place. If the gambler and the computer server running the gambling program are both in a jurisdiction where gambling is legal and the activity is properly licensed, then the activity is legal. But is it still legal when the player is in a jurisdiction where it is not...
legal to gamble but the server is? What if two gamblers are betting in the same poker game, but one is in a jurisdiction where it is legal and the other is not? Is one engaged in an illegal activity but the other not, even though they are playing the same game? What if the players are in a legal gambling jurisdiction but the server is not, or the network passes gambling-related data traffic through computers in a jurisdiction where it is illegal? While we could propose simple and consistent legal interpretations, e.g., that what matters is where the gamblers are, it does not mean that the courts are necessarily free to apply them. They must also weigh issues of public interest, legal precedent, and often, the decisions of other courts, or even the treaties and legal bodies that constitute international law.

The issue of tele-agency and gambling has, in fact, been addressed explicitly, not by a court exactly, but by the dispute-resolving mechanisms of the World Trade Organization (WTO). The WTO is a multiparty international treaty organization whose rules and decisions are binding upon member nations. In 2003, the small island nation of Antigua petitioned the WTO against the United States for their enforcement of antigambling laws on gamblers within the United States who logged in to computer servers in Antigua to engage in gambling [6], [7]. Antigua argued that the actions of the United States in enforcing those laws hurt their ability to engage freely in trade with a market of gamblers in the United States. As the servers were in Antigua, they argued that the gambling was in Antigua, and the United States was engaged in protectionism by denying those players the opportunity to engage in free trade with legal businesses in Antigua. The United States argued that permitting players in the United States to gamble online undermined their ability to use the law to enforce a public moral interest and to maintain social control within its borders. In 2005, the WTO ruled in favor of Antigua.

In accepting this argument, the WTO effectively legalized online gambling in all WTO member nations, provided that gamblers used computer servers located in a jurisdiction that is a member of the WTO and in which gambling was legal. The effective modifier here is significant, because the WTO is not really a court, and a WTO member nation could still choose to enforce their antigambling laws, though they would be subject to WTO penalties and fines for protectionism, or they could withdraw from the WTO altogether. It is also important to note that the basis of this decision is not simply that online gambling is legal, because the servers are in Antigua where gambling is legal. This is unlike other precedents in international law for two important reasons. First, because the WTO does not have legal authority beyond its member nations, and thus, its legal decisions do not carry the weight of precedent that, e.g., the decisions of the International Criminal Court would. Second, because the WTO only has authority over international trade, future use of this decision as a precedent would only be applicable in other cases involving free trade and protectionism among member nations.

That said, we can envision a variety of possible scenarios in which an online activity involved the use of teleoperated robotics and was a matter of free trade. That is, where the activity involved would be illegal if it were engaged in locally, but a commercial industry might exist in which people were willing to pay for the opportunity to circumvent local laws through remote teleoperation and thus the WTO decision would apply as a precedent. For example, in 2004, a Texas entrepreneur launched a Web site (www.live-shot.com) that, for a fee, allowed users to log in, aim, and fire a real gun at real targets. His ultimate plan was to provide live animals for a teleoperated hunting business, claiming that this would serve a market of physically impaired hunting enthusiasts who could not go out into the woods themselves [8], [9]. The business and Web site are now defunct, because 11 states including Texas passed laws making online hunting illegal by requiring the hunter to be physically present when hunting.

Interestingly, the Texas law prohibits anyone hunting with a robot within the boundaries of the state but would not necessarily apply to hunters in Texas going online to, e.g., hunt big game in Africa with a robot. If the laws were written so as to prohibit the act of online hunting itself then the online gambling precedent would apply. If the hunting range was set up in Antigua, for instance, and represented a legal and profitable business interest in Antigua, laws prohibiting online hunting in Texas would be unenforceable due to the WTO ruling.

It is worth noting that the morally abhorrent nature of the activity in the gambling case was not sufficient to justify enforcing the local laws over promoting free international trade [6], though perhaps some activities could reach a level of abhorrence that this would no longer be true. The animal hunting case would not seem to rise to such a level. But we could imagine jurisdictions that either lacked certain legal prohibitions or decided to permit certain activities to generate trade revenue by attracting customers wanting to engage in teleoperated activities, precisely because they are illegal or prohibited in the locales where their online customers reside. Because of this, the commercially successful activities are likely to descend toward the questionable and prurient end of the moral spectrum, including sexual acts, violent acts toward animals and humans, or human degradation and torture. This raises a disturbing set of questions: What if there was a jurisdiction...
willing to sell the opportunity to execute people who have been sentenced to death? or which allowed humans to consent to risking their own lives in mortal combat with teleoperated robots? And, as human slavery remains marginally legal in a handful of countries, would the consent of slave owners be sufficient to decriminalize physical violence up to death against a slave by a robot operator in another country in which such slavery and violence is criminal? In a globalized economy that has already seen banks and multinational corporations establish offices or incorporate in jurisdictions that offer them the greatest protection from taxes or other legal liabilities or restrictions, we should not be surprised when certain locales seek to enrich themselves by becoming safe havens supporting the circumvention of other established legal jurisdictions. Should robotic crimes fall through similar jurisdictional cracks as online gambling and offshore tax havens, we might well see the emergence of some sort of robot safe-havens.

Beyond the economic and trade aspects, there are critical issues of interjurisdictional enforcement as well. Even if it were acknowledged that an illegal act was committed in jurisdiction A, using a robot being controlled by a clearly identified person in jurisdiction B, it is not clear that courts in jurisdiction B would necessarily be able to prosecute the offender. Jurisdiction A would first need the person to be arrested and extradited from jurisdiction B to prosecute them, but not all countries have extradition treaties with each other, and, even where there are treaties, not all crimes warrant arrest or extradition.

A promising legal concept that might serve to prevent the use of robotic technologies to exploit local differences in legal standards, or circumvent prosecution due to jurisdictional gaps, is the universality principle or universal jurisdiction [10]. It was famously used by Spanish courts to arrest and charge Augusto Pinochet for crimes he committed as the dictator of Chile, though he never stood trial. However, the justification for applying the principle of universal jurisdiction is that the crimes committed are so heinous that they are crimes against all of humanity, and thus, all courts have the authority to prosecute suspected offenders. It is doubtful that most specific cases involving robots would rise to the level of crimes against humanity. It is also doubtful that Spanish courts (or the Belgian courts that also assert universal jurisdiction) would have an interest in prosecuting tele-agency crimes around the world or have the resources to do so. We might instead imagine that certain specialized courts could be constituted and supported by international treaty organizations, such as the United Nations, or by perhaps expanding the scope of crimes considered by the International Criminal Court. But such a development would likely arise in the face of public outcry in multiple countries at the inability of existing legal structures to reign in a growing number of such crimes.

Conclusions
In summary, the development and use of teleoperated robotic systems will continue to present new difficulties for the enforcement of local and international laws. These systems present a new capability for committing violent crimes at great distances that did not exist before. Moreover, the ability of tele-agency to separate actors from their actions will further enable the exploitation of inconsistencies between the legal standards of different jurisdictions. These legal issues are likely to be exacerbated by recent developments in international trade and globalization. There are some counterweights to these rather bleak possibilities however. First, robots only provide a margin of anonymity to their controller and not complete anonymity. Second, there are fundamental asymmetries in tele-agency, such that information can be transmitted in both directions, but material entities and properties are stuck on the effector end of the robotic system. Finally, the jurisdiction issues could be addressed by international courts or universal jurisdiction, but the establishment of such courts is unlikely, and most cases of telerobotic crimes will fail to rise to the current high standards set for universal jurisdiction.

References

Peter M. Asaro, Department of Media Studies and Film at the New School University, New York, NY, 10011, USA. E-mail: asarop@newschool.edu.
The meanings of ethical concepts and rules, in a given situation, should be clear and unambiguous. If they are not, one must undertake to clarify their meanings to the extent possible (..) New ethical judgments and cases should be assimilated, where possible, into the existing body of cases, rules, laws, policies, and practices.

Terrell Ward Bynum on Norbert Wiener

Since 2003, the ethical, legal, and societal issues (ELS) in advanced robotics have attracted the increasing and lively interest of academic and professional circles. A similar, although more occasional, debate has also spread to the general public, stimulated either by the novel statements of researchers about recent advancement in robotics or by new and sensitive robotics applications. This increase of contributions and interest occurred hand in hand with the rapid development of research and applications in the service (personal) robot sector, marking the end of the robot segregation era [1].

Many are the current social motivations for a high demand for personal robots. On the economic side, the transitional conclusion of a bull cycle for
industrial robotics; on the social side, the new demand for dependable and safety automatic autonomous machines to be employed in human assistance. Although the so-called social robots have not yet been sold as standardized market products for consumers, we can foresee that the research and application will point in this direction, driven both by the market demand and the challenges and richness of technological and scientific issues posed to the roboticist researchers (see the concept of social caretaking [2]).

Some major segments of our lives already depend on sophisticated machines. However, in some critical instances (where robots are entrusted with human lives, such as in medicine, human care, or war theaters), we are approaching an ethical and regulatory gap, primarily because of the lack of new criteria to ascribe responsibility to software agents and learning machines.

In fact, the recent thrust in research and applications programs aiming toward developing robots that are able to cooperate with humans—that are, software agents and robots entrusted with high learning and decision-making functions—presents many theoretical and practical cases in which moral and legal aspects of the responsibility-attribute problem for learning robots may soon become an urgent issue [3]. The responsibility-assignment problem is a central issue in human–robot interaction (HRI).

These and other epistemic limitations and gaps in our current moral codes and regulatory guidelines, concerning software agents, learning machines, and robots, required, in the view of roboticist Gianmarco Veruggio—who coined the word Roboethics and started the debate on this issue—an effort of analysis and original research in this new field of ethics in science and technology.

Roboethics is the applied ethics developed for 1) the identification and analysis of ethical issues that arise from current and upcoming robotics applications and 2) the possible definition of some general guidelines on the issues mentioned earlier. In a broader meaning of the word, roboethics not only applies to the so-called negative rights (that is, the prohibition of some actions) but also the identification of some trends of robotics research and development that enhance what are viewed as human positive rights (the encouragement to some actions): robotics research and applications that promote and enhance human rights for well-being, education, right of medical care, food, housing, and employment. Actually, robotics and bionics systems may potentially make a significant contribution to the solution of many open human issues. Assistive robots could promote the right of people to live a life of independence and social participation; robotic systems in medicine could foster fundamental human rights by improving the quality of medical operations and protecting the patient’s physical integrity; biorobotics have the potential to provide effective therapeutic means to restore lost motor functionalities; and robots for the environment could be extremely useful in the cleaning of polluted areas.

Following the first International Symposium on Roboethics [5], one of the first steps in roboethics was to select a methodology for the identification and analysis of technoethical issues in robotics. To our knowledge, three European projects (which also availed themselves of international experts as well as background studies in information and computer ethics and bioethics) represent the most structured efforts in this field. These are as follows: the Euron Roboethics Atelier, 2006 [6]; Ethicbots 2006–2008 [7]; and Coordination Action for Robotics in Europe (CARE) 2008 [8]. Here, the analyzers could 1) identify the most universally shared and enforced declarations, conventions, and agreements about human rights; 2) achieve a shared knowledge about the ethical notions involved in the potential violation and promotion of human right in relation to human interaction with robotic, bionic, and artificial intelligence (AI) systems; and 3) apply their contextual understanding of these ethical notions to some technoethical cases. This article has adopted the same methodology.

The further step was to identify the human rights to be protected and promoted by roboethics, listed and discussed in the Universal Declaration of Human Rights (1948) and the Treaty of Lisbon (2000). These are as follows:

- respect and protection of human dignity and privacy
- right for physical and integrity of the person
- right for liberty and security
- right to protection of personal data
- right for the elderly to lead a life of dignity and independence and to participate in the social and cultural life
- right to integration of persons with disabilities
- fair access to technological resources and social and cultural discrimination (per ages, gender, and census).

A further phase resulted in applying roboethics analysis to the following cases:

- Human dignity and privacy: In what way could robotics, in all its applications, affect human functions, capabilities, and rights related to human data protection and privacy issues? On the side of positive rights, how could these technologies suggest a further definition of the concepts of human liberty, dignity, and identity?
- Preservation of human identity and transhumans: In the field of bionics and robotics prosthetics, what is the border between restoration and enhancement? What is the human identity that robotics implants should preserve?
- Liability and responsibility issues: Who is responsible for the possible malfunctions and damages by autonomous robots to humans and/or property? (issues in the domain of AI and law. This subject also implies the identification of new possible additions to the definition of personhood and agentive capacities and the responsibilities discharged regarding robots).
- Psychological effects: The effects of personal robots on human logical and emotional structures and on human relationships (psychology, sociology, HRI, neurosciences, and law).
- Cost-benefit analysis: The comparison between robotics applications and all other possible alternatives, evaluating
whether or not to add the ethical element among the variables listed in the cost-benefit analysis. This subject also comprises the issues of human replaced by robot, job displacement, and analysis of new educational requirements and professional qualification of the human workforce and operators (economy, public policy, engineering economics, and corporate social responsibility).

Initially, the fields of robotics that are going to raise more urgent and entangled ethical issues are biorobotics and robotics military applications. The reason is that these two sectors have, with their research and prototypical applications, immediately and directly intervened in the core of the existing corpus of ethical principles, regulations, and laws related to the most sensitive issues of human life. In less dramatic but not less-important instances, assistive and educational robotics has also given rise to some ethical concerns in the field of HRI.

Because of the rich and complex debate in roboethics and the far-reaching scenarios that could develop over the following decades, this article (with the partnership of the aforementioned European projects) adopted a triaging work methodology, analyzing the issues that had the following elements:

- **Novelty**: Issues that have never been coped with; the absentia legis and the lack of regulations, in many cases (bionics and military robotics), underpin a severe responsibility gap.
- **Emerging**: Issues arising in a nonprogrammed way, as the prototypical robotics products are the results of the drive of different instances: research and business.
- **Complexity**: Issues lying in the crossroads of several disciplines (robotics, AI, moral philosophy, psychology, anthropology, and law).
- **Social pervasiveness**: Issues related to current and yet-to-be-released robotics products.

### Ethics in Complex Technological Societies

For the purpose of this study, we will provide the various meanings of the term *ethics* within the context of modern debate. Traditionally, ethics is the philosophical or theological subject that studies human behavior and assessment criteria for human behaviors and choices. Modern ethics have developed from various points of view along classical philosophy. In the last decades, in our highly complex societies characterized by technoscientific development, the attribution of moral and practical responsibility is becoming more and more difficult with regard to unintentional or collateral consequences of actions and operations that are produced by unidentifiable decisions (group decisions, complex administrative structure, distribution of responsibilities, and computerized operations). Often, one finds it impossible to attribute final responsibility to a single person or to a defined social entity. In the absence of a definition and a precise analysis of the responsibility chain, technologically advanced societies have shifted the issue onto the concept of risk assessment, thereby attributing value to the damage produced by entities that are seemingly devoid of responsibility. In our case, the question is: Who is responsible for any damage that may be caused by an autonomous robot? Is it the designer, manufacturer, programmer, or final user? Often, it will be difficult to obtain easy answers to this question.

From the viewpoint of ethical theories, we observe that individuals possess the ethics of common sense, which provide them with the moral guidelines for decision making, from the small ones to important life decisions. This affects the day-to-day life as well as our actions in our close-knit social relations. Individuals often adopt different ethical theories from time to time: utilitarianists for some decisions, in other situations, or moments generally go by the terms of their upbringing, customs, tradition, and religion (descriptive ethics). These seem to be enough for them.

However, this line of thought reaches its limits when, as social figures or in our profession, we are faced with complex problems, in which our actions may have multiple consequences that are difficult to follow and predict; and when our common sense is faced with the problems we’ve never dealt with before—for example, bioethical dilemmas. In these situations, the ethics of common sense leads to various paradoxes: we find ourselves without any conceptual resources and in the difficult position of having to pass judgment. In these cases, we need a logical-critical set of ethics (critical ethics) that 1) reveals the implicit and, perhaps, never uncovered assumptions in our ethics of common sense and in our outdated ethical theoretical base and 2) analyses the reasons, the pros and cons, and their origin. Inevitably—we may not even realize it—we resort to prescriptive ethics and ethical theories.

In practice, when faced with general and complex topics, we may refer to the fundamental, relevant values involved in our dilemmas; we may adhere to the more updated morality applied to issues close to ours; or we try to step up from the universally shared prescriptions toward new ethical frontiers. Prescriptive ethical theory, which develops and justifies the principles of moral actions, refers to ethical theories and related guidelines. These in turn represent the general ideas that contain ethical principles to reach an internal and systematic coherence. In defense of the ethical principle, the ethical philosophers either implicitly or explicitly refer to an ethical theory.

In the 20th century, because of the dissatisfaction raised by the issues that are deemed to be the limits to traditional ethical theory—utilitarianism and deontologism inspired by Kant—ethics became fragmented into several forms: rights ethics, virtue ethics, feminist ethics, and applied ethics [9].

For instance, according to virtue ethics—which have less to do with single actions, but more to do with different life styles and ways of life—the most important moral obligation is our personal relationship with the action of making or using robots and the analysis of the effect of robots with respect to the concept of good and happiness: Which robots could contribute most to the happiness to the full and complete quality of life of human beings? [10].


another instance, considering the need to account for values that are part of a collective conscience, the so-called rights ethics sees human rights to be the most relevant elements in common for a system of cultural and moral pluralism and as being reasonably deemed to be the final manifestation of universal ethics.

One of the challenges to traditional ethical theories by some thinkers regarded the limitation of moral considerations to the problems of humans and their relationship in human society, regardless of other living organisms and environment. Therefore, ethical theories and the new applied ethical theories, in particular, were called upon to include nonhuman entities in their analysis (animal ethics, environmental ethics, and planetary ethics known as the Gaia theory) as well as human products (bioculture, computer ethics, and roboethics).

The various applied ethical theories are, in turn, connected and intertwined with other disciplines, including law, sociology (descriptive ethics), economics, and various scientific fields. One of the central themes of applied ethics is the concept of responsibility, which is a moral notion. Legal responsibility determines the rules of the relevant prescriptive ethics, meaning the group of commands and prohibitions adopted by a society or group and that also defines professional ethics. For the purpose of this article (to analyze the human–robot relationship), we shall consider the two main meanings of the term responsibility: 1) the analysis of the identity of the agent of the cause of certain actions and their effects (utilitarianism or consequentialism or teleologism) 2) an expression of motivations that leads an agent to act in a certain way (deontological ethics or Kantian ethics), according to which the individual assesses the consequences of his or her actions.

In the last century, we know that the questions “What authority and what set of moral rules am I obliged to be accountable to? State law? God?” forced many answers (utilitarian or deontological) to a crisis point. Often, the moral response of the individual as well as the moral evolution of our society have led to an opposition to the responsibility toward the nation, church, or traditional roles of social institutions (see Max Weber’s ethics of intentions or ethics of individual conscience).

Some serious events related to World War II changed the notion of responsibility and differentiation of roles (i.e., engineers deal with engineering, doctors with medicine, soldiers obey orders from superiors, etc.) and also some famous legal cases, e.g., the Adolph Eichmann trial, during which he defended himself by stating that he was obeying the orders under his administrative responsibilities and that he did not decide on, nor was he aware, of the entire project of Jewish extermination or the debate that ensued after Hiroshima and Nagasaki 1945.

Furthermore, there are different cases in which our personal moral clashes with those adopted and imposed by the society we live in and whose law we live under. Such is the case, for example, of the death penalty (which is in force in various states in the United States, although contested by internal currents), animal rights, abortion, or euthanasia. In these cases, the implicit moral philosophy in state laws does not coincide with the feelings of many groups of the relevant societies, which leads to this sort of conflict.

Therefore, we observe that, in contemporary societies, the notion of responsibility is not limited to the moral consideration of actions of the agent or cause and deals with the needed conformity of the action with a group of duties (therefore, the analysis of consequences is less decisive). In today’s society, just the elements of complexity and technological ruling determines the aspect known as heterogenesis of ends, according to which our actions may have consequences that are extremely difficult to estimate, which may even be opposite to our intentions. According to Morin’s ecology of action [11], once the action departs from the individual, it lives a life of its own, and combines itself with the environmental conditions (social models, actions, and reactions of other agents), and the final result is beyond the agent’s predictive abilities.

Faced with this vacancy in the attribution of individual responsibility, some researchers attempted to identify collective shared responsibilities, were it to be impossible or vain to identify an individual responsibility. In the case of scientific research, science and technology studies (S&TS) expert René von Schomberg proposes to adopt an assessment system based on foresight and knowledge (foresight and knowledge assessment). The author sustains that, because the definition of responsibility is considerably more arduous to define in scientific fields, due to the unintentional consequences, uncertainty, or ignorance of results, instead of identifying the ethical responsibilities post hoc, it is necessary to establish the ethics of the overlap of knowledge between different areas beforehand (synergy: scientists, politicians, etc.), because the quality of knowledge shall determine the ethical value of the applications that will follow.

At the same time, one must constantly ensure that maximum precision of predictions to identify both the wholesomeness of research and relevant applications as well as the potential ethical problems [12].

Other authors have emphasized the need to avoid the overlap between ethical problems and technical solutions: among the latter, the expert of computer ethics, Abbe Mowshowitz [22] states that the seemingly eternal social problems are real enough, but to look for their cause in technique or autonomous technology is both mistaken and harmful. We should not blame technology for human failures. (…) Autonomous technology contributes to the belief in technological determinism, i.e., reinforces belief in the inability of people to make significant choices in their lives. It directs attention away from wielders of power to systems of reified collectivities. The law is smarter than the social sciences—it defines the corporation, for example, as a fictional person for purposes
of assigning responsibility and does not absolve key actors of their responsibilities. Institutions should be seen as convenient fictions that help explain individual decision and action (..) Only the actions of human beings can be alienating or dehumanizing. Reification of technology allows for an illegitimate transference of responsibility from persons to a fictional social construct, and at the same time, impedes our ability to come to grips with the very real ethical challenges posed by the uses of technology.

In conclusion, we observe that, in roboethics, the definition of moral responsibility and the resulting notion of liability—which is central in human–robot relationship—could differ according to the philosophical assumptions which, knowingly or unknowingly, have been adopted.

Roboethics or Robot Ethics?
Roboethics was originally conceived as human-responsibility ethics. The roboticists and ethicists that contributed to their creation highlighted the following aspects:
- problems regarding robot autonomy
- problems related to warfare applications
- problems in human–robot relations (dependence, privacy, robot appearance, and potential confusion between natural and artificial)
- digital divide (for nations, genders, and ages)
- ethical dimension of technology.

According to Veruggio [13], roboethics is an applied ethics whose objective is to develop scientific/cultural/technical tools that can be shared by different social groups and beliefs. These tools aim to promote and encourage the development of robotics for the advancement of human society and individuals and to help preventing its misuse against humankind.

This and similar definitions imply that robotics and its applications are subject to moral judgment and human intervention. According to this perspective, roboethics is not artificial ethics and not even the regulatory system of dependability and safety. Roboethics indicated that the individual and society may intervene upstream on the direction of robotics and its products. The individual may, for example, limit the use of robots, according to a wider precautionary principle, in the absence of the necessary safety precautions when missed for humans.

As researchers, individuals may refuse to design robots that are deemed harmful or hazardous; as professional bodies, they may discuss and decide upon an appropriate professional ethic.

As already discussed, roboethics has adopted the principles established by the Charter of Human Rights and the Lisbon Treaty. However, these certainly do not satisfy the realm or depth of the ethical debate. If last century extended the ethical realm to an increasing number of elements, including animals and our planet among those entitled to rights, we witness an increasing application rights ethics to even broader categories, thereby crossing over the human/organic barrier.

In the definition of roboethics, as developed by some authors [6], [14], the agreement with the thesis of American moral and political philosopher John Rawl and to his reflective equilibrium is explicit, according to which the charters and treaties, although advanced in an attempt to associate the highest number of participants and commendable in an attempt to widen the promotion of rights to a greater number of members and functions as possible, do not satisfy the possibility of positive progress, nor can they provide for the entire range of criminal uses of robots.

During the past six years of discussion on roboethics, various positions have emerged regarding the ELS issues in robotics. Just as some authors have continued to use the term roboethics, others have used robot ethics. These two terms don’t always indicate different notions. However, robot ethics has been often used to indicate 1) the artificial ethics of robots or the morality of robots and 2) the group of prescriptions, rules, and regulations regarding robot safety and dependability. Although the term robot ethics may be used in the latter sense, because roboethics also studies issues of dependability and safety, the difference between roboethics and robot ethics is more complex.

Some authors see robot ethics as an artificial morality and the robots as autonomous agents. Their theses in this context are different, but they can be linked to this cluster of arguments.

The first states that sophisticated autonomous robots, because they are equipped with intelligence and a certain kind of conscience, they include an ethical system that can learn and evolve or are able to decide between good or evil. According to this position, morality and immorality constitute a gradual continuum: such is the case for children, which, in our society, are not deemed to be fully responsible for their actions or the disabled [15], [16]. As they are quasi-moral agents, robots are also subject to ethical behavior, equipped with some rights of their own.

The second position in robotics morality states that robots, as devoid of emotion and passion, and as they are equipped with rational behavior (because that is how they are programmed), could be more ethical than human beings [17]. These two positions have one aspect in common: they state that, when faced with the complexity of robot technology, it is difficult to assign responsibility for any possible damage they may cause; or when faced with irrational and complex situations (war theaters and other circumstances where an immediate answer—such as a gut reaction—is required), the placing of trust in an automated morality is the only solution. Thus, the focus has shifted from the issue of human ethics and has moved onto operational artificial ethics, which therefore avoids the debate begun by Sanremo [5] and that was essentially focused on the issue: when is it right to limit the autonomy of robots as final products?

When one tries to shift an ethical problem onto a technical solution, there are some concepts that fall into...
the background, such as 1) important underlying assumptions about the adopted ethical theories; 2) the relevant glossary, meaning an ambiguous use of terminology. In our case, here the keyword is machine learning.

Although searching for new solutions is more than laudable and auspicious—for instances, when faced with the horror of war and above all those of this century—coherence with the philosophy of law imposes that every new form of ethics cannot demean the rights acquired by the previous ones: “Any new ethics must deal with the same substance as the old role responsibility ethics, namely, with values and norms that restrict or delimit human action and thus enable or guide traditional decision making” [12]. This statement means that roboethics and robot ethics should be coherent with the shared sets of rules.

Overall, every thesis in favor of robotics morality per se is based on the assumption of predictability (or rationality and foreseeability) of the behavior of an autonomous robot. A robotic morality provides, in the conception of its authors, that the robot shall behave ethically according to what it has been programmed with and what it has learned. In a word, it cannot deviate from its incorporated laws: unlike a human being that, although supposedly brought up to be good, may decide to do evil unexpectedly, a robot will always decide ethically because it cannot behave otherwise.

Here, the critical analysis should focus on two critical elements: 1) the kind of ethics that the robot expresses and 2) the unpredictability issue in leaning machines. The first question implies, in our view, that no robot ethics can avoid the needed, deep and broad debate on the human ethical principles grounding roboethics and their enforcement. The second critical element is that, according various experts [18], [19], “programmers, manufacturers, and users may not be in the position to predict what a learning robot will do in normal operating environments and to select an appropriate course of action on the basis of this prediction” [3]. This means that not only it would be advisable that learning robots should be self-evidently distinguishable from nonlearning robots, but that in the former case, the learning process be made transparent for the robot user.

An ethically accurate analysis of the bonds and limits of the potential sphere of action of a robotics morality should indeed take into careful account the epistemic and logic-hidden faults related to both ethical theories and technical constraints. The fact that the current development of non-segregated robotic systems does not yet allow for a precise modeling of every possible environmental factor or for a final definition of normalcy in rich operating conditions should be the basis for a more prudent discourse on intelligent robots behaving ethically.

Moreover, a careful assessment of the responsibility issues in attributing to robot rights, which belong only to humans, and the analysis of the problems encountered in assessing the environmental factors affecting robotics behavior, would advise to adopt the enlarged version of the precautionary principle, limiting robotics autonomy when all the environmental factors are not precisely assessed.

Conclusions: An Open Debate

From the manifold and depth of the considerations around roboethics over the last six years and considering the new ethical issues involved that have never been broached, it appears as if, beyond the different point of views, one major side question is posed, which would need a more general answer: Which direction robotics is going to take? And which should it take? The hope of the many, whose consideration of the importance of robots in society, goes hand in hand with ethical related concerns, is that a promising alliance between robotics and the field of science and technology studies should also be swiftly established.

It has been widely recognized—although not very often practically accepted—that institutional practices in science and technology are tacitly shaped and framed by deeper social values and interests [20]. These include

- important political–economic relationship to science and technology
- shifting from science as independent republic to science as cooperating in innovation and in the knowledge economy
- impacts of the increasing commercialization of science in particular areas affecting public trust
- loss of credibility and senses of unease in the general public [21].

The researches in science and society developed by social scientist Sheila Jasanoff and collaborators have sustained that technoscientific knowledge stabilizes in society through a complex and articulated process of negotiation and then very seldom experts’ opinions are exempt from uncovered assumptions.

It should be clear to all the parties involved in the process of molding roboethics that this free, open, and relatively untroubled debate is possible, because until now no dramatic incident in the field occurred. And, we hope it never will. Robotics research is driven by future scientific visions, and it so should be. Limitations of the freedom of scientific research should be very carefully discussed and almost never imposed. Limitations should be painstakingly decided on for marketed robotics applications.

However, some from the robotics industry have already expressed their concern that, at the first sense of unease from some social groups toward new robotics applications, some limits will be decided on. Increasingly, if a dramatic incident was to occur, the extraordinary character of the situation will impose severe constraints. We have already witnessed those occurrences in bioethics (stem cells, etc.). Among the recommendations resulting from the analysis of ELS issues in robotics developed in the frame of CARE, there is a need to “avoid that ELS issues in robotics could become a barrier to further progress of our field.”

If some incidents were to occur, the outdoor process of discussion will be channeled in the indoor environment of the experts committee, relegating the other parties (roboticists,
ethicists, stakeholders, and society as a whole) to the role of concerned observers.

For these reasons, we feel that all people concerned with roboethics should take vantage from this fortunate window of fresh and free debate, defining with careful consideration and wise temperance of language the general ethical assessments and rules for future robotics.

References


*Fiorella Operto*, Scuola di Robotica, P.O. Box 4124, P.za Monastero 4, 16149 Genova, Italy. E-mail: operto@scuoladirobotica.it.
Approaches to programming ethical behavior for computer systems face challenges that are both technical and philosophical in nature. In response, an incrementalist account of machine ethics is developed: a successive adaptation of programmed constraints to new, morally relevant abilities in computers. This approach allows progress under conditions of limited knowledge in both ethics and computer systems engineering and suggests reasons that we can circumvent broader philosophical questions about computer intelligence and autonomy.
Incrementalism is the view that progress toward a goal is made in a stepwise fashion; it is thought to be applicable, especially, in circumstances in which it is difficult to know at the onset what will be the proper and efficient means that will allow the goal to be reached [11]. Uncertainty about the means typically owes to the complexity of the problem to be solved, and incrementalism becomes more attractive as the need for some progress toward the goal (however, piecemeal) becomes more pressing. Machine ethics, in the sense in which I will here discuss it, is the goal of properly constraining computer-controlled machines (robots and other multifunction computer systems), where the grounds for the constraints are the ethical reasons. Since machine ethics as an academic discipline is in its infancy, there are basic philosophical questions surrounding its plausibility—ones that concern the nature of moral agency and responsibility for non-human actors—it is perhaps unwise to wait for a complete philosophical account of its objectives and methods. For instance, in the definition of machine ethics mentioned earlier, we could easily get sidetracked by questions over whether the ethical reasons need to be unanimously endorsed by humans or whether they would count as reasons for the computer.

Basic issues in philosophy, as any historian of the discipline will attest, take a long time even to formulate correctly, and some of them may never be resolved. When these two notions are combined, incrementalism in machine ethics becomes a practical proposal about how to simultaneously engineer and provide an ethical sanction for the kinds of information technologies that are taking over the many activities—once performed solely by humans—that are generally assumed to have moral relevance. In fact, technological societies have already traveled some distance down the path of replacing human action with technology. We now have technologies that assist us to fight wars, keep public order, monitor air and water quality, provide medical care, execute financial transactions, distribute electrical power, and so on. These activities have moral relevance because human and animal lives and welfare depend on them. Many people living in contemporary societies are dependent on machines for their well being, and they will likely become more dependent as machines gain in their functionality and are deployed in further domains. So while we are already getting the machines (like it or not), we desperately need the ethics that ought to accompany them. This is why I call machine ethics a pressing problem.

Versions of incrementalism have already been developed in academic studies of social choice under conditions of bounded rationality [18], of the federal budgeting process [21], and in international relations [1]. Since the 1960s, incrementalism has been at the center of many debates in the social sciences. Incrementalists, in general, share a criticism of the synoptic or rationalist-comprehensive view of public administration, political science, and social choice. The starting point for this criticism is the recognition that, in trying to solve complex social problems, humans suffer from cognitive limitations and the decentralization and wide distribution of mechanisms for decision making [6]. The rationalist expectation that policy scientists must start out already knowing how to proceed often leads to wasted theorizing or a paralysis of indecision. The central study with which incrementalism is most often identified came from former Research and Development Corporation researcher Charles Lindblom’s work on the social science of public administration that began with his article “The Science of ‘Muddling Through’” [12] and continued in a series of widely discussed publications, chief among them being The Intelligence of Democracy [13]. To account for partisanship in public policy decision making, Lindblom developed with the philosopher David Braybrooke the theory of disjointed incrementalism [4]. Lindblom’s classic account included the notion of adaptive incremental adjustment to respond to a prior decision in the strategic decision space [11]. Though incrementalism has met with resistance from rationalist economists and political scientists [3], [15], Lindblom’s account of muddling through has become one of the most widely read pieces in the social sciences [16].

Incrementalism has now emerged in implicit forms in the recent literature on machine ethics. Wallach and Allen [20] profess an interest in the incremental steps arising from present technologies that suggest a need for ethical decision-making capabilities and explore the prospects for a bottom-up approach to building an artificial moral agent (AMA) that has many of the features of an incrementalist approach. The notion of adding ethical constraints as a machine takes on new functions is connected to Johnson and Powers’ account of the role responsibility of computers as surrogate agents [9]. But the roboticist Ron Arkin has come closest to advocating an explicitly incrementalist position. Much of Arkin’s work concerns lethal autonomous robots in the context of warfare. In describing the step-by-step development of a fieldable ethical system for such robots, Arkin writes that “we can initially represent a small set of forbidden and obligated constraints and test the overall system without the necessity of a fully complete set of representational constraints” [2]. He thinks that, given the impending widespread introduction of lethal warfare robots, “baby steps are better than no steps toward enforcing ethical behavior in autonomous system warfare.”

Abstracting from these studies in the machine ethics literature, we can explore in greater depth what it means to be committed to incrementalism in machine ethics, both as a description of the (very young) practice of applying ethical constraints to machines and as a normative model for the development of machine ethics for highly integrated and safety critical computer systems, such as those for warfare, air traffic control, and public health. The
supposed benefit of thinking about progress toward a goal from an incrementalist perspective is that doing so will avoid the often-paralyzing search for a final theory when circumstances are not ripe for understanding what that theory might look like. The contrasting synoptic view of machine ethics suggests that we could conceive of a machine ethic in its totality, prior to knowing what capabilities the machine will have or the time frame during which they will be developed. Incrementalists, on the other hand, need not start out with the assumption that they know where they are going; they just want to decide (given the circumstances): what’s the next step? I start off, then, with several basic suppositions. There is a pressing need for progress in machine ethics, and we have no final theory of ethics, nor any good idea of how the steps of a machine ethical theory might develop.

In what follows, I will try to explain and evaluate incrementalism for machine ethics. Borrowing a term from decision theorists [11], I will sketch a particular version that I call adaptive incrementalism in machine ethics (AIME), and describe how it might address both practical and philosophical problems that have already become apparent in the literature. In the final section, I will look at the potential criticisms of AIME, especially the suggestion that a machine ethic that is developed in a piecemeal fashion must remain incomplete. (This criticism can be summed up thus: “you can’t get there from here—incrementally or otherwise”) I will then introduce an argument from limited behaviorism that circumvents the criticism that no machine can be said to act ethically without our first having established that it is a free moral agent. The argument from limited behaviorism, I will urge, does not reduce to a general behaviorism for human ethics.

Limited behaviorism does, however, provide a response to an objection to machine ethics that is based on an objection to machine intelligence—an objection made famous by John Searle’s Chinese Room argument. The Searlean objection to machine ethics is that a machine could never become ethical because it cannot develop—either incrementally or otherwise—to a point at which it would become conscious. What the objections to machine intelligence and machine ethics share are doubts about a computer’s ability to have intentional states (including states like “intending to act” and “preferring X over Y”). Machine ethics must also face the longstanding supposition that having and acting because of intentional states are necessary conditions for something being a moral agent. While the objection to machine intelligence is too complex to address in the space available, the objection to machine ethics must be met, for if it is correct, then it will be just as senseless to talk of machine ethics as to talk of machine social psychology.

Adaptive Incrementalism
As a model to develop machine ethics, incrementalism operates on two levels: as adaptations to the commands or program of the machine (software level) that act as ethical constraints and as additions to the capabilities of the hardware/software (system level) that trigger new ethical constraints. Let us begin with the latter first.

Candiates for the kinds of machines that will need ethical constraints have a hierarchy of system capabilities. While we need not order these capabilities, we generally know that the stored program architecture, for instance, is more basic than the ability to implement a voice recognition program, or to play a game of chess, or launch a rocket. Some approximate division can be made between the basic system capabilities and the morally relevant ones. As a start, let us say that

1) The computer system gains morally relevant capabilities as soon as some human being could be made worse off by the designed action of the system.

2) The computer system gains morally relevant capabilities as soon as some human being could be made worse off (and no one better off) by the designed action of the system.

This formulation eliminates the concern for accidental harms that might come, for instance, from a CPU falling on someone’s foot or causing an electrical fire because of a faulty wire. Still, it casts a very wide net for morally relevant capabilities and opens up the door for disputes about moral theory. Surely, we will have to look closely at the different ways in which someone could be made worse off, and whether these kinds of harms count as morally relevant. As a first refinement, let us adopt Pareto’s criterion of efficiency and say that
programmed ethical constraint must have one of the three abstract descriptions:
3) Allow the machine to act on the capability unchanged.
4) Meliorate the way in which the human being might be made worse off.
5) Disallow the capability entirely.

As an example, consider the addition of the ability to transfer files to and from another networked computer through the standard file transfer protocol (FTP). One ethical constraint would be to add a capability to lock certain files in the target computer, thereby disallowing reading or writing to the files. Another constraint would be the introduction of password-protected access such as secure file transfer protocol (SFTP).

The previous example is almost trivial in its ethical import. Consider, however, the additional capabilities that networked computers gained when suddenly they could attach viruses to e-mail or hack into a target computer and replace system files. The addition of these capabilities already has called forth adaptations: firewalls and antivirus software. This is an example of AIME, in practice, and one that comes fairly early in the young history of machine ethics. It is interesting that such changes were broadly adopted without large public debates about the correct moral theory to use. Viruses were deemed bad, no matter what moral theory one held.

The serial implementation of constraints of the sort in 3)–5) would constitute the development of an adaptive machine ethic. At another level of abstraction, the list of ethical constraints for some machine at a certain point in time might read as follows:
6) Protect the privacy of clients.
7) Protect the property rights of clients.
8) Maintain the health of bystanders.

At a more fine-grained level, the designers of the ethical constraints might describe 6) as prevent unauthorized access to files containing medical data of client and similarly for each ethical constraint we could have more clearly operationalized commands for the machine to follow. This feature of differing descriptions of constraints, based on the level of abstraction, is also apparent in human ethics. People never directly act on the constraint don’t be bad. Rather, someone’s action might be governed by the constraint don’t cause needless harm to people or even more concretely, don’t strike dan now.

An important feature of AIME is that there is no a priori list of ethical constraints for a machine; each constraint is developed because of, or in response to, an additional capability. This means that there are no necessary components of an adaptive incremental ethic. Also, there is no limitation, in principle, on the entities that will gain moral status and thus come under consideration when thinking about who (or what) is potentially made worse off by the computer’s new capability. In the formulation given earlier, I included only human beings in 1) and 2), but a machine ethic could develop to include animals, ecosystems, future generations of humans, etc.

To most ethicists, the potential conflict of values in 6)–8) (privacy, property, and health) will be readily apparent. If there is a software adaption that presents a choice between these three values, then there will be no theory-independent way of ranking the outcomes. The problem of conflicting values is not so acute in situations where one and the same machine receives, serially, the adaptations in 6)–8); this merely represents a pluralism of values. Still, what happens when we must choose only one adaptation—when, for instance, we must choose to protect someone’s health at the expense of someone else’s informational privacy?

Lindblom’s account of muddling through addresses this very problem, albeit in the context of the conflicting values that every public administrator must consider. He states the problem thus:

How does one state even to himself the relative importance of these partially conflicting values? A simple ranking of them is not enough; one needs ideally to know how much of one value is worth sacrificing for some of another value. [16]

and provides the following answer:
The value problem is . . . always a problem of adjustment at a margin . . . . That one value is preferred to another in one decision situation does not mean that it will be preferred in another decision situation in which it can be had only at great sacrifice of another value [12].

The key to the incrementalist’s sanguine acceptance of value conflict is that he gives up on a rational-comprehensive account of the system as a whole. Lindblom admits that “[e]xcept for roughly and vaguely, I know of no way to describe—or even to understand—what my relative evaluations are for, say, freedom and security, speed and accuracy in governmental decisions” [12]. That very same uncertainty about relative evaluations might hold if we replace governmental decisions in Lindblom’s statement with information technology decisions. For AIME, this means that ethically adapted robots and computer systems will betray the multiple ethical perspectives of their designers, and indeed, this multiplicity will be unattractive to proponents of a rational-comprehensive view of machine ethics.

For the incrementalist, though, appeal to a rational-comprehensive view is not useful for actual ethical decisions, because in the context of decision making, it is not a view available to him or her. Computer systems, like other technologies, often evolve over time frames that encompass different users and designers (indeed, sometimes, even different teams of designers); they are affected by laws and regulatory policy, markets, available infrastructure,
and changes thereto [8]. No designer has the option to select one instantiation of the system to last forever. All changes are temporary in principle.

Following Lindblom's account of incrementalism, we would describe each choice to adopt an ethical constraint as an instance of a successive limited comparison. The relevant question for this comparison is, does the designer prefer the new system—with both its additional, morally relevant ability and its programmed ethical constraint—to the system as it was? Of course, there will always be the possibility that the designer’s preference will diverge from some users’ preferences, but the fact is that a choice must be made. We should not assume that a designer’s preferences are independent of those of the users nor that they are equivalent. In any event, the user is not in a position (typically) to make design decisions, and the designer’s knowledge is limited by many factors.

AIME thus represents a description and a normative proposal for design of machines from engineering and ethical perspectives. We may engineer certain capabilities into a machine and find, at a later date, that there was a better way to achieve the same results. Likewise, we may put in place a certain ethical constraint for a machine, to adapt it to society, given its increased functionality. At some later date, we may find that this is the wrong constraint or that there is a more precise way to constrain the machine—allowing it to do more or do less than we originally allowed it to do. Incremental adaption therefore suggests an ongoing process to address new machine capabilities and to reevaluate old constraints in light of the new capabilities.

Incrementalism, as applied to the problem of machine ethics, should have the same advantages as those promised for the original incrementalism—the science of muddling through in public policy. In his study of incrementalism for policy making, Hayes identifies five virtues of incrementalism (summarized here): 1) facilitates action where the rational ideal is paralyzed 2) reduces the costs of analysis by providing a defensible basis for confining attention to some alternatives over others 3) facilitates learning from mistakes 4) facilitates majority building my minimizing disruption to established practices 5) the failure of any given step to solve a particular problem often makes the best case for taking the next step.

Advocates of incrementalism claim to have confirmed these characteristics in empirical studies in several areas of the social sciences, while critics have countered with other studies. As Knott et al. point out, “[t]he concept of incremental adaption entered the social sciences literature because empirical observations of behavior did not fit with a fully rational approach to decision making” [11]. So even if some studies did validate the rational-comprehensive model, the advocates of incrementalism were able to argue that that model works in far too few circumstances to be adopted. In addition, they were able to show that the logic of incrementalism under conditions of cognitive limitations (bounded rationality) promised more success for decisions under those circumstances [14]. Much of the force behind incrementalism as a movement in the social sciences thus depended on both a priori and a posteriori arguments for it, as its advocates were able to shift adeptly between descriptive and normative conceptions of the model.

Criticisms of AIME

It is safe to say that theorists and practitioners of machine ethics—a diverse group consisting mostly of philosophers, ethicists, computer engineers and programmers, and artificial intelligence (AI) enthusiasts—are predisposed to rationalist accounts. Hence, it is unlikely that a muddling through approach to machine ethics would be adopted tout court. We will now turn to consider some of the challenges for the AIME version of incrementalism.

Since adaptive ethical constraints in AIME are triggered by system-level changes to machine capabilities, a question of scope will arise even before the first constraint is operationalized. The question is, are all capabilities of machines relevant? My earlier refinement of the definition of a trigger for morally relevant system changes attempted to forestall this worry. It is likely, though, that some designers will initially be stumped and may consider too many capabilities as relevant. The other extreme, of course, is to see the trigger as operating hardly at all. This latter outcome, I believe, is our current situation with computer system development. Without ethicists participating in design, almost all system-level abilities are taken to be morally irrelevant. (Arkin, a roboticist, is one of the few who takes moral relevance seriously) To overcome this potential problem with AIME, designers and ethicists will have to work jointly toward a proper moral sensitivity for the computer systems they are designing. This will require a good deal of imagination—an appreciation for plausible what if? scenarios—and also timely feedback about their systems in the testing and initial deployment phases.

Conceivably, the ethicists and designers may fail to develop the same level of sensitivity, and this could lead to disagreements and conflicts. The second major criticism, then, questions whether it is really plausible in AIME to expect people from different backgrounds and with disparate objectives to
come together to make steady progress. If this isn’t possible, then AIME will suffer the same fate as the incrementalists found in rationalist-comprehensive practice: paralysis. One way to allay the worry about internecine disputes among teams of designer ethicists is to impose on the team the same set of institutional interests. For private sector projects, for instance, a company might tie compensation or promotion to the total performance of the robot or computer system. That is, they would be wise to consider a successful system as one that meets simultaneously the criteria of high functioning and ethical propriety.

Making AIME a corporate benchmark may not entirely solve the problem, however. Some computer systems could be of a scale that they would pit the interests of one corporation against another or even one nation against another. This will most likely be the case with warfare robots. For instance, the primary ethical constraint for some military commanders will be to minimize friendly fire casualties. Of course, the system may be programmed also to respect the laws of war, minimize collateral damage, protect noncombatants, and observe proportionality of destructive force. As is clear from Arkin’s prototype implementation [2], not every objective can be pursued at once. How these objectives are weighted may vary between nations and even between computer systems.

One feature of AIME that may make some (but certainly not all) ethicists uncomfortable is the lack of theoretical unity of ethical constraints that are developed over time. This concern also affects the definition of the system trigger for ethical constraint. But the heterogeneity of these machine ethics may go deeper. So far, we have spoken merely of ethical constraints, but (allowing for developments in software) it will likely happen that complex machines in the far future will also need ethical rules, perfect and imperfect duties (in the Kantian sense), and maybe even a conception of virtues. This eventuality is not troubling for the incrementalist, since the motivating assumption is that machine ethics starts out by doing what it can do and not worrying about what it cannot do in imparting ethics to machines. It is likely that, over time, the programmed ethical constraints of a particular system will be superseded by future constraints. It is even possible that entire schemes of machine ethics that were once considered successful will be found at some point to be inferior to new schemes. This mimics one of the most interesting aspects of the history of the electronic digital computer. Hardware configurations, storage media, input/output, and programming languages undergo revolutions of sorts (though not all simultaneously) when improvements are developed. We should expect the same for machine ethics.

**Ethics, Consciousness, and Agency**

The very notion that machines could take on ethical abilities faces philosophical challenges from many quarters, most of which we cannot consider here. Comparing machines to typical (human) moral agents, philosophers have insisted that machines lack free will, consciousness, and morally relevant emotions such as regret, empathy, and shame. Perhaps the most succinct complaint against attributing excessive abilities to computers comes in John Searle’s attack on the notion that computers can think [17]. This suggests an objection to machine ethics by means of a rather obvious extension of Searle’s argument against computer thought.

Searle denies that computers have anything more than syntactical abilities in operating according to their programs. He thinks that semantic ability—something only minds have, as far as we know—is the key to conscious understanding. In addition to lacking semantic ability, computers lack the ability to have their own intentional states, on his view. Certainly, they can represent intentional entities—for instance, with sentences of a natural language displayed on a screen. And they can follow syntactic rules to display new sentences, as his Chinese room argument suggests. This is a mere simulation of thinking, according to Searle; it is not the real thing. Simulation, in the sense of supplying answers to queries, isn’t sufficient for thinking, understanding, or consciousness, according to Searle. The case against machine intelligence, for Searle, is simply open-and-shut.

When it comes to the free will of a computer, here Searle is somewhat less emphatic. He concedes that if “somebody built a robot that we became convinced had consciousness, in the same sense that we do, then it would at least be an open question whether or not that robot had freedom of the will” [17]. Nonetheless, he is quite sure that this will not happen, as he thinks the computer would have to have the abilities of a human brain in order for it to have consciousness and intentional states of its own.

Why does Searle’s syntax/semantics argument against AI not doom AIME? First, we must realize that consciousness in machines is not a necessary condition of their ethical behavior. It may well be a necessary condition for a machine to be self-aware, and hence aware of its intentional states—should it have them—as being the free cause of its actions. But at this point, we should acknowledge a point that Searle takes to be in his favor, but that actually cuts against his arguments. We poorly understand the physical and neurological bases of consciousness and intentionality that the emphatic argument against machine intelligence overreaches what we know. When we do come to understand the brain better, it may well turn out that intentionality and consciousness is possible for computers. Searle’s syntax/semantics argument is just too simple to rule out that possibility.
Second, we might understand intentional states in computers as merely those states that are about, or represent, or are directed at states of the world through the models that they use. As the computer scientist Brian Cantwell Smith wrote, “there is no computation without representation!” [19]. When I look at a radar animation, from a computer system, of storms gathering over Delaware, those images and the representations that generated them are about the weather outside. This weaker sense of intentionality allows that computers have intentional states—in fact, they must have them insofar as they have representations that connect to models of the world. Computers do not (now) generate their own intentional states, independent of the representations that we program them to have. But as Smith points out, we put representations in computers as merely those states that are about, or represent, or are directed at states of the world through the models that connect to models of the world. Computers do not (now) generate their own intentional states, independent of the representations that we program them to have. But as Smith points out, we put representations in computers—as representations that serve as a partial model of the world—when we design them. The having of representations about the world is sufficient for computers’ having intentional states in the weak sense.

If we attribute weak intentional states to computer systems and suppose that AIME is successful in constraining the behavior of the system—for ethical reasons—we have the essentials of a system that could be seen to be behave ethically in a world that it represents in it programs. Having arrived at the heart of the issue, this behavioral definition of machine ethics, I will argue, is a genuine beginning to machine ethics.

To see how this is so, consider what I will call the causal–rational tradition in ethical theory from modern and contemporary philosophy—one exemplified in the works of Donald Davidson [5]. According to this tradition, when we act ethically, we act for (or with) a certain reason, but not any old action for a reason will suffice. An outwardly right action done for the wrong reason is not an ethical action, but neither is it necessarily unethical. For instance, paying a debt because you fear retribution or saving a child’s life because you thought it was your son (but you were mistaken) are taken to be ethically neutral on this view. Right act, wrong reason.

If not all (outwardly appearing) right acts are ethical acts, and the ethical acts are determined by the reasons that caused them, what determines what the right acts are? This has been a challenge that the causal–rational tradition has yet to answer. Let us sketch a way around this problem.

Suppose some (not inconceivable) future in which all of humanity converges on one theory of ethics. Call it theory T. We settle all disputes about which things have moral status, what we owe to them, etc. To perform all of the right actions, we know all of the self and other regarding obligations, we know the correct virtues and how to practice them, we know which preferences count, and we know how much they count. That is, we know all of the ethical reasons for all of the right acts, and we humans are now able to do the right acts for the ethical reasons. Consider building a machine that fulfilled all of the right acts of theory T because it was programmed through AIME. (By this time, our moral trigger would have become much more sophisticated.) For instance, if T consists of obligations and permissions, the machine acts on all of the obligations in the appropriate contexts and never commits an impermissible act. What would we have to say then about the ethics of that machine?

The opponent of AIME might claim that the machine didn’t do the right acts (as determined by our complete theory T) for the right reasons, and in this, she would be correct, but only because the machine didn’t do the acts for any reasons at all. Machines, the opponent insists, aren’t capable of acting for a reason. The proponent of AIME should, I think, accept this criticism. Moreover, this very fine result—that the machine does exactly what T prescribes—is all we could want anyway, given our current state of knowledge about computing machines and ethics. This is a result that I call limited behaviorism: a machine behaves ethically in doing all and only the right acts. This kind of behaviorism does not offer a replacement for a comprehensive ethical theory for humans, but it allows that, in principle, we may develop a machine that performs all of the right acts (and none of the wrong ones) that we would expect of any human operating according to theory T. That is, limited behaviorism is not reductive on the issue of ethical reasons for humans; a human acting for ethical reasons is still defined separately from the right acts that the human performs. But limited behaviorism does accept the equivalence of right acts, whether they issue from a machine or a human. It also finds that, since there are no machines that act for ethical reasons, there will be no practical difference between a right-acting machine and an ethical machine. In the event that machines develop to the point where, one day, they can have reasons for acting, they too would have to distinguish their (outwardly) right acts from their ethical acts.

Finally, the AIME opponent might raise a different kind of objection: that we’re nowhere close to figuring out what theory T is, so if we go about designing and programming a machine with our current imperfect state of moral knowledge, we are bound to impart some mistakes to the machine’s ethical build. Compare now our current attitude toward the inculcation and practice of ethics for humans, in this (quite imperfect) state of moral knowledge in which we find ourselves. Most of us are quite sure—on both theoretical and practical levels—that we aren’t close to discovering theory T. Is that a reason for not teaching ethics.

One feature of AIME that may make some ethicists uncomfortable is the lack of theoretical unity of ethical constraints that are developed over time.
to our children? Is that fact a reason to give up, ourselves, on trying to do what is right? I think the answer to these questions is clearly no.

The propriety of muddling through when it comes to machine ethics is reinforced by an analogy to the theory of moral development in the cognitive/structural account of moral education that was introduced by the Swiss psychologist Jean Piaget and is now widely associated with Lawrence Kohlberg’s theory of the stages of moral development [10]. Wallach and Allen address Kohlberg’s theory explicitly as a useful way of thinking about machine ethics [20]. The basic idea in Kohlberg’s work is that no child is born with its full complement of moral abilities, and some never develop to the highest stage of moral development, but there is progress in teaching a child to reach a higher stage. Indeed, after Kohlberg’s theory became well known, many scholars came to doubt whether he had correctly described and ordered the stages. But as far as I know, no serious scholar ever proposed that parents and teachers cease moral education of children until the theorists could come up with one comprehensive account of both moral development and moral knowledge.

In the spirit of Kohlberg, roboticists might set themselves the task of constantly refining the ethical abilities of their machines. Yet, it is unclear that there is an ultimate stage of ethical behavior to be reached. Ethical performance may be relative to the kinds of tasks to be performed, and the moral complexity of the environment in which the machine operates. If this is so, the requirements of machine ethics may always get more demanding. Human parents might hope to teach their children to reflect on the nature of that ultimate stage—something that is likely to be exceedingly difficult to program into a machine.

This analogy between parenting and programming returns us to the main distinction in the beginning of this article: that incrementalism is a means for the goal of machine ethics. Incrementalism does not purport to list the stages in the development of machine ethics, and our current poor state of moral knowledge does not provide a clear-enough picture of what a right acting machine will look like, even on the assumption that it can be built. Thus, ethicists and designers find themselves in a situation like that of new parents. They have responsibility for a new, developing being, but aren’t sure how best to effect its moral education. They themselves lack complete moral knowledge—they await (but despair of never discovering) a perfect theory T of morality. They would like to be able to impart practical reasoning to this new being but realize that this is a job for later years, and—first things first—they have to get the child simply to behave. They are faced with a daunting task, but one that is pressing due to the likelihood that doing it imperfectly, by muddling through, is bound to be better than not doing it at all.

**Acknowledgments**

This work was supported in part by the U.S. National Science Foundation under grant EPS-081425.

**References**


**Thomas M. Powers**, University of Delaware, Newark, DE, 19716, USA. tpowers@udel.edu.
This is the first installment of a two-part tutorial. The goal of the first part is to give the reader a basic understanding of the technical issues and types of approaches in solving the basic path-planning or obstacle-avoidance problem. The second installment will cover more advanced issues, including feedback, differential constraints, and uncertainty. Note that this is a brief tutorial rather than a comprehensive survey of methods. For the latter, consult some of the recent textbooks [4], [9].

Motion planning involves getting a robot to automatically determine how to move while avoiding collisions with obstacles. Its original formulation, called the piano mover’s problem, is imagined as determining how to move a complicated piece of furniture through a cluttered house. Have you ever argued about how to move a sofa up a stairwell? It has been clear for several decades that getting robots to reason geometrically about their environments and synthesize such plans is a fundamental difficulty that recurs all over robotics.

The stages of motion-planning development are parallel to those of an integral calculus: 1) The integration problem was clearly identified and defined; 2) perfect, exact solutions were developed for many classes of functions; and 3) since these were limited to a small subset of functions that people care about, numerical integration methods were developed with great success in practice. The similar stages of motion planning were as follows: 1) it was clearly defined in the 1970s; 2) the 1980s saw the development of perfect, combinatorial solutions, which are ideal in some settings, but not practical in most; and 3) the 1990s brought sampling-based methods that are not as elegant but offer practical solutions to modern industrial-grade problems. Over the past decade, motion-planning algorithms have been widely used in robotics and automation and have furthermore found applications well beyond, including the fields of virtual prototyping and computational biology.

**Problem Formulation**

Let \( \mathcal{W} \) denote the world that contains a robot and obstacles. For a two-dimensional (2-D) world, \( \mathcal{W} = \mathbb{R}^2 \) and \( \mathcal{O} \subset \mathcal{W} \) is the obstacle region, which has a piecewise-linear
The basic path-planning problem is informally summarized as follows: given an initial placement of the robot, compute how to gradually move it into a desired goal placement so that it never touches the obstacle region. See Figures 1 and 2 for examples.

Consider the task in terms of algorithm inputs and outputs.

- **Inputs:** An initial placement of the robot, a desired goal placement, and a geometric description of the robot and obstacle region.
- **Outputs:** A precise description of how to move the robot gradually from its initial placement to the goal placement while never touching the obstacle region.

The output description will be a path through a set of all intermediate transformations of the robot from start to finish.

**Living in C-Space**

Although the motion-planning problem is described in the world, it really lives in another space: the set of all rigid-body transformations that can be applied to the robot is called the configuration space or C-space. Finding a solution leads to computing a path through the part of the C-space that avoids robot-obstacle collisions.

A rigid body may translate and rotate. Most people are much more familiar with performing one transformation to place a body into a scene rather than thinking about all transformations. The notion of configuration space was the key insight to Lagrangian mechanics of rigid bodies [1], as it allowed dynamics to be expressed using the precise degrees of freedom of a body. The idea was introduced to motion planning by Lozano-Perez [12] and Udupa [17].

The C-space in physics and control theory is usually called a Lie (pronounced Lee) group. In this context, which is much more widely studied than motion planning, the C-space is considered as a differentiable manifold, which leads to considerable technical and notational hurdles. The C-space used in motion planning requires no calculus; therefore, it is described as a topological manifold, which is fortunately much simpler to define and manipulate. The definition of an n-dimensional (topological) manifold $C$ is a subset of $\mathbb{R}^m$ for $n \leq m$, such that every $q \in C$ is contained in at least one open subset of $C$ (pick a small one) that is homeomorphic. (Homeomorphic means that for an open set, say $O$, there exists a continuous, bijective function $f : O \rightarrow \mathbb{R}^n$ for which the inverse $f^{-1}$ is also continuous to $\mathbb{R}^n$.) The intuition is that, in the local vicinity of every $q$, a manifold behaves like $\mathbb{R}^n$. It is a nicely behaved surface. The existence of sharp corners does not even matter;

![Figure 1.](image1.png)

**Figure 1.** A 2-D example of basic path planning.

![Figure 2.](image2.png)

**Figure 2.** A 3-D automotive assembly task that involves inserting or removing a windshield wiper motor from a car body cavity. This problem was solved for clients using the path-planning software of Kineo CAM.
However, branching or the locally changing dimensions is not allowed (Figure 3).

We now take a look at the C-spaces that commonly arise in planning. Consider a 2-D world. Let \( A \subseteq \mathbb{R}^2 \) denote a polygonal robot. It could, for example, be all points inside of a triangle defined by vertices \((-1,0), (1,0), \text{and } (0,1)\). We could rotate the robot counterclockwise by any \( \theta \in [0, 2\pi) \) and then translate it by any \( x_1 \in \mathbb{R} \) in the \( x \) direction and any \( x_2 \in \mathbb{R} \) in the \( y \) direction. This allows for any possible position and orientation, and every \( x_1, y_1, \theta \) combination leads to a unique robot placement. Let \( q = (x_1, y_1, \theta) \) be called the configuration. A point \((x,y) \in A\) would then appear at some \((x',y') \in W\) (in the world) given by

\[
\begin{pmatrix}
  x' \\
  y'
\end{pmatrix}
= \begin{pmatrix}
  \cos \theta & -\sin \theta & x_1 \\
  \sin \theta & \cos \theta & y_1 \\
  0 & 0 & 1
\end{pmatrix}
\begin{pmatrix}
  x \\
  y \\
  1
\end{pmatrix},
\]

(1)

which uses a standard 3 by 3 homogeneous transformation matrix. The upper left \( 2 \times 2 \) block is just a rotation matrix.

The set of all configurations \( q = (x_1, y_1, \theta) \) is clearly a subset of \( \mathbb{R}^3 \), but to define the C-space, we must take into account that \( \theta \pm 2\pi \) yields equivalent rotations. We write that \( C = \mathbb{R}^2 \times S^1 \), in which \( S^1 \) denotes a circle in the topological sense and accounts for \( \theta \) (the circle is obtained by gluing \( 0 \) and \( \pi \) together). The C-space \( C \) is a 3-D manifold, and each element is nicely described as \( q = (x_1, y_1, \theta) \). Remembering that \( \theta \) wraps around at \( 2\pi \) is crucial to motion planning; otherwise, an artificial barrier or redundant exploration will be introduced. If the robot is not allowed to rotate, then we obtain the translation-only case and \( C = \mathbb{R}^2 \) with \( q = (x_1, y_1) \).

For the 3-D world, the concepts mostly extend as you might expect. Three translation parameters \( x_1, y_1, z_1 \) appear, and a translation-only robot then has a C-space \( C = \mathbb{R}^3 \) with \( q = (x_1, y_1, z_1) \). However, the set of 3-D rotations turns out to be 3-D manifold all by itself, and it is not as simple as a circle or sphere topologically. The best way to see its structure is to use quaternions to represent rotations. Since this a brief tutorial, only the essence is given here, and quaternion algebra is avoided here as it is not technical term for the resulting space is real projective three space, denoted \( \mathbb{R}P^3 \). For the case of a 3-D robot that can translate or rotate, we obtain \( C = \mathbb{R}^3 \times \mathbb{R}P^3 \), which is a six-dimensional manifold. We can represent the configuration as \((x_1, y_1, z_1, a, b, c, d)\) while enforcing that \( a^2 + b^2 + c^2 + d^2 = 1 \). The use of quaternions means that the set of all 3 by 3 rotation matrices is parameterized by \( a, b, c, \) and \( d \):

\[
\begin{pmatrix}
2(a^2 + b^2) - 1 & 2(bc - ad) & 2(bd + ac) \\
2(bc + ad) & 2(a^2 + c^2) - 1 & 2(cd - ab) \\
2(bd - ac) & 2(cd + ab) & 2(a^2 + d^2) - 1
\end{pmatrix}.
\]

(2)

With different possible parameterizations of rotations, for 2-D or 3-D worlds, it is important to realize that if two points are close under one representation, they might be far under another. Furthermore, if there are singularities in the parameterization mapping (e.g., yaw–pitch–roll representation), the C-space might not even represent the same manifold as the set of all rotations.

Now that different possibilities for \( C \) have been presented, consider the parts of \( C \) that are prohibited due to collision. Let \( A(q) \subseteq W \) denote a closed set of points in the world occupied by the robot \( A \) when it transformed to configuration \( q \). A configuration \( q \in C \) places the robot into collision if and only if \( A(q) \cap O = \emptyset \) (the robot and obstacle are attempting to occupy at least one common point in \( W \)). The set of all noncolliding configurations is often called the free space and is defined as

\[
C_{\text{free}} = \{ q \in C | A(q) \cap O = \emptyset \}.
\]

(3)

The complement is called the obstacle region in C-space:

\[
C_{\text{obs}} = C / C_{\text{free}}.
\]

The problem statement given in the “Problem Formulation” section seemed somewhat informal; however, using the C-space, the basic path-planning problem can be precisely defined: given a robot description \( A \), an obstacle description \( O \), a C-space \( C \), an initial configuration \( q_1 \in C \), and a goal configuration \( q_G \), compute a continuous path \( \tau : [0, 1] \rightarrow C_{\text{free}} \) with \( \tau(0) = q_1 \) and \( \tau(1) = q_G \) (Figure 4). A

**Figure 3.** The first three are manifold, because they locally look like \( \mathbb{R}^2 \); the last two are not because at some points the dimension changes or branching occurs.
a typical way to express $\tau$ is a sequence of line segments, which ignores the particular parameter $s \in [0, 1]$, but is good enough for motion-planning results. Note that the path must be continuous; otherwise, the robot would appear to teleport from one place to another, which is obviously cheating. Gradual motions through $C$ make the robot move gradually through $W$.

**Combinatorial Planning**

Although the motion-planning problem is in the continuous $C$-space, its computation is discrete. Therefore, if we want an algorithmic solution, we need a way to discretize the problem. This has led to two main schools of thought: 1) combinatorial planning, which thrived in the 1980s, constructs structures in the $C$-space that discretely and completely capture all information needed to perform planning and 2) sampling-based planning, developed mainly across the 1990s, uses collision-detection algorithms to probe and incrementally search the $C$-space for a solution rather than completely characterizing all of the $C_{\text{free}}$ structure. The second approach is most widely used in practice; however, the first one is far superior in many instances. Therefore, it is worth to study both.

To illustrate the philosophy of combinatorial planning, consider the case in which $W = \mathbb{R}^2$ and contains a point robot ($A = \{(0, 0)\}$) that cannot rotate. In this case, $C = \mathbb{R}^2$, and the task is simply to connect the dots in the plane with a curve that avoids the obstacles [Figure 5(a)].

Here is a simple technique that contains all the essential ingredients of combinatorial planning. All the methods first compute a road map, which is a graph in which each vertex is a configuration in $C_{\text{free}}$, and each edge is a simple path through $C_{\text{free}}$ that connects a pair of vertices. Here is one way to achieve this:

1) Decompose $C_{\text{free}}$ into trapezoids with vertical side segments. Figure 5(b) shows the result. From each polygon vertex, an attempt is made to shoot rays upward and downward. Each ray may be immediately blocked, or it may travel until hitting another part of the obstacle boundary.

2) Place one vertex in the interior of every trapezoid. It doesn’t really matter where; for simplicity, pick the centroid.

3) Place one vertex in every vertical segment. The resulting vertices are shown in Figure 5(c).

4) Connect each segment vertex to the two vertices that are in the interior of the neighboring trapezoids. Each connection forms an edge in the graph and corresponds to a straight-line path.

The result is a road map that appears to capture the structure of $C_{\text{free}}$. How would you implement these steps? For the first step, we could iterate over each vertex and

---

**Figure 4.** In the C-space, the problem looks simple: connect $q_i$ to $q_G$ while remaining in $C_{\text{free}}$.

**Figure 5.** A combinatorial planning illustration: a) 2-D polygonal obstacle region with proposed $q_i$ and $q_G$ (one possible solution is shown in a dashed path); b) the trapezoidal decomposition; c) constructing a graph by placing a vertex in every vertical edge segment and every trapezoid interior; and d) connecting $q_i$ and $q_G$ to the graph and searching for a solution path.
determine precisely where each upward and downward ray intersects other segments. We could then easily identify the first segment hit by the vertical ray in the above and below directions. For an example as simple as Figure 5(a), this is a fine method. However, if there are \( n \) polygonal edges in total and \( n \) is large (say, \( n = 20,000 \)), then the method is not efficient because it takes time \( O(n^2) \).

By proceeding carefully, this computation can be reduced to time \( O(n \log n) \) by employing the plane sweep principle [6], which underlies many decomposition algorithms used for combinatorial planning. First, sort the polygon vertices from left to right, requiring time \( O(n \log n) \). During the algorithm execution, a list of some polygon segments is maintained and sorted from top to bottom, as they are stabbed by a vertical line. The method proceeds incrementally from vertex to vertex, traveling from left to right. At each step, the edge list is updated by simple insertions and deletions, which each take \( O(\log n) \) time using self-balancing binary search trees. If the edges incident to the vertex are both to the left, then the two edges are deleted from the list. If they are both to the right, they are inserted into the list (in order). Otherwise, the one to the left is deleted, and the one to the right is inserted. Thanks to this ordering, and we can determine in \( O(\log n) \) time the segments directly above and below the vertex, which are first stabbed by upward and downward rays. It is furthermore simple and efficient to incrementally extend the graph as each vertex is processed. For more details, see Section 6.2.2 of [9] or Section 6.1 of [6].

The road map is constructed without considering the query pair \( q_1 \) and \( q_G \). Once the investment is made, the same road map can be used for multiple query pairs. In other words, we can easily solve numerous motion-planning problems in a world that contains the same obstacle and robot. Here is a simple way to use the computed road map from Figure 5: 1) find the trapezoids that contain \( q_1 \) and \( q_G \), 2) connect \( q_1 \) and \( q_G \) to the vertices in their respective trapezoids, 3) search the graph for a path that connects \( q_1 \) to \( q_G \).

The first step can be performed trivially in \( O(n) \) time by testing whether \( q_1 \) (or \( q_G \)) lies in each trapezoid; this can be shaved down to \( O(\log n) \) time by developing clever hierarchical point-location data structures [6]. The second step takes constant time, and the final step can be performed in \( O(n) \) time using simple graph search algorithms such as breadth first or depth first.

For the simple case of a point robot in a polygonal world, numerous alternative algorithms exist that yield comparable performance. We could, for example, decompose \( C_{\text{free}} \) into triangles instead of trapezoids. The general principles are that each cell should be easy to traverse (convex is ideal), the decomposition into cells should be easily computable, and the adjacencies between cells should be straightforward to determine. Based on these properties, a useful road map is obtained.

Road maps need not be obtained by cell decompositions. For example, a shortest path road map yields distance-optimal paths and is constructed by connecting certain pairs of vertices that can see each other, and each has an interior angle greater than \( \pi \). A maximum clearance road map can also be computed efficiently. In general, a road map is expected to have two properties to be useful for planning:

1) Accessibility: It is simple to reach a point on the road map from any \( q \in C_{\text{free}} \) while trivially avoiding collisions.
2) Connectivity preserving: For any pair \( q_1, q_2 \) of points that is connected to the road map, a path exists between them in the road map if and only if there was a path between \( q_1 \) and \( q_2 \). In other words, if \( q_2 \) is generally reachable from \( q_1 \), then traveling between them via the road map must also be possible.

It seems up to this point that combinatorial planning solutions have beautiful properties. Most importantly, they construct a discrete representation of the problem that exactly captures the solution. In other words, there are no approximation or sampling errors. These methods are called complete, meaning that, for any input problem, they correctly determine in finite time whether or not a solution exists.

Here comes the trouble. Most motion-planning problems involve robots that are not modeled as points and they can rotate in addition to translating. How many of these nice combinatorial planning ideas extend? First, consider the case of a polygonal translation-only robot. If the robot \( A \) and obstacle \( O \) are convex polygons, then \( C_{\text{obs}} \) is a polygon in which every edge corresponds to a point-to-edge contact between \( A \) and \( O \). See Figures 6 and 7. Can you see how to achieve this by reassembling the edges of \( A \) and \( O \) into \( C_{\text{obs}} \), with the edges appearing in an ordering with the edge normals? Once this conversion is made, a trapezoidal decomposition approach is easily applied. If \( A \) and \( O \) are nonconvex, then they need to be first

![Figure 6. A triangular robot and a rectangular obstacle.](image)

![Figure 7. (a) Slide the robot around the obstacle while keeping](image)
decomposed into convex pieces to construct the convex pieces of $C_{\text{obs}}$. A trapezoidal decomposition algorithm could even be used for the convex decomposition of $A$ and $O$.

Now introduce rotation. For the translation-only case, $C_{\text{free}}$ has a piecewise linear boundary because the translation is a linear transformation. Unfortunately, the rotation is nonlinear and commonly represented using trigonometric functions. Various ways to reparameterize rotation matrices lead to improvements; however, nonlinearity is unavoidable. For computation, polynomial parametrizations are preferred. The previous piecewise-linear representations are then replaced with semialgebraic representations, meaning that each facet of $A$, $O$, and $C_{\text{obs}}$ is represented as the roots of implicit polynomials. Constructing $C_{\text{obs}}$ in terms of polynomial roots is straightforward, but a combinatorial explosion occurs that produces far too many facets for practice (the example in Figure 6 already produces more than 70). For 3-D problems, it becomes considerably worse. The next difficulty is to perform cell decomposition. The first motion-planning method to accomplish this is the cylindrical decomposition method of Schwartz and Sharir [13], which produces a number of cells that is doubly exponential in the dimension of $C$. More efficient cell decomposition methods exist, and there is Canny’s algorithm [3], which directly produces a road map through $C_{\text{free}}$ in a singly exponential time without a prior decomposition. These methods provide solutions to the general path-planning problem; however, they are even rarely implemented due to numerical issues and inefficiency from the combinatorial explosion.

### Sampling-Based Planning

Sampling-based approaches are by far the most common choice for industrial-grade problems, because $C_{\text{obs}}$ is composed of an unwieldy number of facets. They abandon the idea of explicitly characterizing $C_{\text{free}}$ and $C_{\text{obs}}$ and essentially leave the planning algorithm in the dark when exploring $C_{\text{free}}$. The only light is provided by a collision-detection algorithm, which is a black box that probes $C$ to determine whether some configuration (or a small ball around it) lies in $C_{\text{free}}$. These algorithms often work by hierarchically representing $A$ and $O$ and attempting to quickly determine collision at a course resolution [11]. Many collision detection methods are incremental, which means that they can yield extremely fast performance by saving information from a previous execution on a nearby configuration.

Planning algorithms then work by incrementally probing and searching $C_{\text{free}}$ for a path, gradually revealing more and more of it with the collision detector. In this way, motion planning feels like using a robot with a weak sensor to explore an unknown environment. This might seem odd since $O$ and $A$ are given; however, the environment being explored is $C_{\text{free}}$ (or equivalently, $C_{\text{obs}}$), which is high dimensional and prohibitive to explicitly represent. Sampling-based approaches attempt to find a solution quickly while cheating their way out of building a full map of $C_{\text{free}}$. Don’t compute more than you have to.

To get a feeling for sampling-based planning issues, we first introduce a frequently used method based on rapidly exploring random trees (RRTs). Figures 8 and 9 show the algorithm and its result. The idea is to aggressively probe and explore the C-space by expanding incrementally from an initial configuration $q_0$. The explored territory is marked by a tree rooted at $q_0$. Each iteration extends the tree by adding a leaf vertex and edge that connects it to the rest of the tree. Each edge is a collision-free path between two configurations. The RRT algorithm picks a point $q_{\text{rand}}$ at random in $C$ (not $C_{\text{free}}$) and then tries to connect the tree to it by extending the nearest point in the tree. This biases the tree toward aggressively reaching unexplored parts of $C$, but eventually settling on uniform coverage.

Some implementation details are needed to clarify Figure 8. Step 1 initializes $G$ to contain a single vertex, corresponding to $q_0$ and no edges. In Step 3, a random configuration generator is used to obtain $q_{\text{rand}} \in C$. A random translation could be selected uniformly from a bounded region (often an axis-aligned rectangle). A random 2-D rotation is easily obtained by randomly selecting some $\theta \in [0, 2\pi)$. It turns out that selecting a uniformly random 3-D rotation is technically more challenging. Here is an amazingly simple method. Choose three points $u_1, u_2, u_3 \in [0, 1]$ uniformly at random and then let [14]:

$$a = \sqrt{1 - u_1^2} \sin 2\pi u_2 \quad b = \sqrt{1 - u_1^2} \cos 2\pi u_2$$

$$c = \sqrt{u_3^2} \sin 2\pi u_3 \quad d = \sqrt{u_3^2} \cos 2\pi u_3$$

in the rotation matrix (2).
What does uniform random really mean for $C$? Recall from the “Problem Formulation” section that the set of transformations could be expressed in numerous ways, meaning that the notion of uniform randomness appears to be arbitrary. There is, however, a well-defined notion of uniformity based on Haar measure, which is beyond this tutorial; see Section 5.2 of [9]. Intuitively, if we rotate the coordinate frame on which the rotations are defined, then the uniformity should be preserved. The methods for rotation above, including (4), achieve this.

Step 4 finds $q_{\text{near}}$, the closest point in $G$ to $q_{\text{rand}}$ (see Figure 10). What does it mean to be closest? This again depends precisely on how $C$ is represented and implies that a distance function has been defined. The distance function $\rho : C \times C \rightarrow [0, \infty)$ is formally called metric and usually satisfies the following axioms for all $p, q, r \in C$: 1) $\rho(p, q) \geq 0$, 2) $\rho(p, q) = 0$ if and only if $p = q$, 3) $\rho(p, q) = \rho(q, p)$, and 4) $\rho(p, q) + \rho(q, r) \geq \rho(p, r)$. In virtually all sampling-based planning algorithms, performance depends on the choice of the metric. It is sometimes difficult to set the relative weights between rotational distances and translational distances (see Figure 11).

Now that the closest has been established, which points in $G$ are checked for being the nearest to $q_{\text{rand}}$? The simplest is check the vertices and report the nearest one. But the closest point among all those explored could lie along an edge. Rather than incurring an expensive computational cost, a common tradeoff is to check some intermediate points at regular intervals along an edge (Figure 12). This introduces an unfortunate parameter to tune but often simplifies implementations (it is also reasonable to avoid all of this and just use the vertices).

Finally, Step 5 extends the tree. If $C_{\text{obs}}$ were empty, then an edge can be made from $q_{\text{near}}$ to $q_{\text{rand}}$. If $q_{\text{near}}$ is a vertex in $G$, then the endpoints of the new edge are $q_{\text{near}}$ and $q_{\text{rand}}$. If $q_{\text{near}}$ is a point along the interior of an edge, then that edge must first be split, with $q_{\text{near}}$ introduced as an intermediate vertex. Since $C_{\text{obs}}$ is usually not empty, there are two issues: 1) A collision-detection algorithm makes sure that we can travel from $q_{\text{near}}$ toward $q_{\text{rand}}$ while staying in $C_{\text{free}}$, and 2) we might not be able to reach $q_{\text{rand}}$ without hitting $C_{\text{obs}}$. If it is not possible to reach $q_{\text{rand}}$, then the new vertex is instead placed at the configuration $q_i$ that gets as close as possible, as shown in Figure 13. (If no progress is possible, then no new edge and vertex are created.)

The RRT algorithm presented in Figure 8 aggressively explores $C_{\text{free}}$; however, if the tree is grown from $q_i$, there is no consideration of $q_G$. Now consider ways to solve the basic path-planning problem using RRTs.

Here is a simple adaptation. Start the RRT with $q_0 = q_i$, and at every 100th iteration, force $q_{\text{rand}} := q_G$ instead of choosing a random configuration. If $q_G$ is reached, then a path has been found from $q_i$ to $q_G$, which solves the problem. This induces a gentle bias toward the goal. At one extreme, we could pick $q_G$ every time, making a beeline for $q_G$. This would fail miserably when an obstacle is reached. Figure 14(a) shows an example in which this would occur. Aggressively attempting to reach $q_G$ by setting $q_{\text{rand}} := q_G$ in every other iteration would still work, but might waste too much effort running into $C_{\text{obs}}$ instead of exploring. Therefore, a light bias, such as every 100th iteration is recommended.

For many problems, though, such a simple strategy is not enough. Figure 14(b) shows a kind of bug trap from which it is difficult to escape. Because of the existence of

---

Figure 10. A new edge is added that connects from the random sample $q_{\text{rand}}$ to the nearest point in $S$, which is the vertex $q_n$.

Figure 11. Rotation versus translation domination: (a) The task is to move the C shape to the right. Rotation dominates. Performance should improve if rotation is weighted heavily in the metric. (b) In this case, the translation dominates and should therefore be weighted more heavily if this fact is known in advance.

Figure 12. For ease of implementation, intermediate vertices can be inserted to avoid checking for the closest points along line segments. The tradeoff is that the number of vertices is increased dramatically.

Figure 13. If there is an obstacle, the edge travels up to the obstacle boundary, as far as allowed by the collision-detection algorithm.
such situations, which commonly occur in practice, a bi-directional search is more effective and popular. The algorithm grows two RRTs: 1) \( G_I \) rooted at \( q_I \) and 2) \( G_G \) rooted at \( q_G \). Instead of always extending the trees using random configurations, half of the time is spent trying to extend each tree toward the newest vertex of the other tree. The following four iterations are repeated:

1) generate \( q_{\text{rand}} \) and use it to extend \( G_I \), obtaining a new leaf vertex \( q_{\text{new}} \)
2) force \( q_{\text{rand}} := q_{\text{new}} \) and use it to extend \( G_G \)
3) generate a new \( q_{\text{rand}} \) and use it to extend \( G_G \), obtaining a new leaf vertex \( q_{\text{new}} \)
4) force \( q_{\text{rand}} := q_{\text{new}} \) and use it to extend \( G_I \).

Steps 1 and 3 are identical to the execution in Figure 8, but for \( G_I \) and \( G_G \), respectively. Steps 2 and 4 trick the RRT by using the most recent vertex from the other tree as a replacement for \( q_{\text{rand}} \). If either of these two steps ever succeed in connecting the trees to each other, then the problem is solved. This method is quite effective for most practical problems, as aggressive exploration from \( q_I \) and \( q_G \) is balanced with trying to connect the trees to solve the problem.

An example that was solved in 2002 by the bidirectional RRT is the famous Alpha 1.0 puzzle introduced by Nancy Amato and Boris Yamrom. The task is to pull apart the twisted nails, leading to an extremely narrow corridor in \( C_{\text{free}} \) through which the solution path must travel. The solution is illustrated in Figure 15. Most problems are not this challenging, and solutions are often found in a fraction of a second. Nevertheless, there are limitations to the method as well as any sampling-based method. It is not hard to construct pathological examples that cause the algorithm to converge too slowly. In some cases, problem-specific heuristics can then be developed to recover performance.

The RRT-based methods fall into a larger family of methods called \textit{incremental sampling and searching}, in which a graph is incrementally constructed inside of \( C_{\text{free}} \). Each method has a vertex selection method, which determines where to expand next from among vertices in the graph. After that, a local planning method constructs an edge from the selected vertex, thereby extending the tree. In the case of an RRT, the vertex selection method picks the vertex closest to \( q_{\text{rand}} \). The local planning method attempts to connect the vertex to \( q_{\text{rand}} \). As an example of an alternative incremental sampling and searching method, the expansive space planner (ESP) [7] selects a vertex with probability that is inversely proportional to the number of other vertices within a ball of predetermined size. The local planning method then connects to a random configuration within the ball, but only with a probability that is inversely proportional to the number of vertices that lie within a ball centered on the random configuration. Another example that falls into this family is the randomized potential field planner [2], which implements gradient descent in \( C_{\text{free}} \) and uses random walks to escape local minima.

A common nuisance with sampling-based planning methods is that the produced paths are jagged as they traverse \( C_{\text{free}} \). This makes the solution animation jumpy; Making the robots to follow such awkward paths is a comically bad idea. Therefore, path smoothing is usually performed to clean up solution paths. Fortunately, it is straightforward to produce a cleaner path once a jagged solution is given. A simple method is to iteratively pick a pair of points at random along the path and attempt to replace the path portion between them with a straight line in \( C_{\text{free}} \). If this survives the collision-detection verification step, then use the linear segment and discard the original part portion. After several dozen iterations, the path is usually much improved.

The discussion so far has focused only on single-query algorithms, meaning that only one \( q_I, q_G \) pair will be given so that there are no advantages of extensive precomputation. Recall from the “Combinatorial Planning” section that planning problems can be quickly solved once a nice road map has been computed that offers the accessibility and connectivity-preserving properties. This motivates a multiple-query approach to sampling-based planning known as a \textit{probabilistic road map} [8]. In this case, a bunch (e.g., 1, 000) of random
configurations are chosen upfront and declared to be road map vertices. Road map edges are formed by attempting to connect each configuration to all vertices within some specified radius (Figure 16). If a road map can be constructed that satisfies accessibility and connectivity preservation with high probability, then it can be used to efficiently search for solutions to multiple initial-goal query pairs. One difficulty is that the road map may have as many edges and vertices as a high-dimensional grid [10], which provides motivation for pruning strategies that attempt to keep the good road map properties while reducing its size substantially. See, for example, the visibility road map variant [15].

To conclude, we should emphasize that a tradeoff has been made by going to sampling-based methods. Recall from the “Combinatorial Planning” section that combinatorial planning leads to complete algorithms: They always find a solution if it exists; otherwise, they report failure. Since sampling-based methods solve problems without fully characterizing \( C_{\text{obs}} \), completeness is reduced to weaker forms. The goal is to ensure that the sampling eventually covers all of \( C \). This can be expressed in terms of dispersion, which is the radius of the largest empty (unsampled) ball in \( C \). Sampling-based approaches usually achieve resolution completeness, meaning that they will find a solution if one exists, but may run forever if one does not, or probabilistic completeness, meaning that the probability tends to one that a solution is found if one exists (otherwise, it may still run forever). For example, the RRT approaches described above lead to probabilistic completeness, partly because the dispersion is reduced to zero with probability one. Resolution completeness can be obtained by replacing the random configuration generator by a deterministic point sequence that leads to zero dispersion in \( C \) in the limit (for example, consider a multiresolution grid that refines forever).

The best way to learn more about sampling-based motion planning is to experiment with the implementations. You could download and install a free library, such as the Open Motion Planning Library from Rice University, the Motion Strategy Library from the University of Illinois, or the Motion Planning Kit from Stanford. If you instead want to start from the basics, then at least downloading a collision-detection package, such as PQP from the University of North Carolina, is recommended.

**Direct Extensions**

Now that the core motion-planning ideas have been explained for the case of rigid 2-D or 3-D robots among fixed obstacles, several straightforward extensions can be covered for which the planning methods are virtually the same.

The formulation given in the “Problem Formulation” section allowed only one moving rigid body. This limited the C-space to having no more than dimension three for \( \mathcal{W} = \mathbb{R}^2 \) and six for \( \mathcal{W} = \mathbb{R}^3 \). If we allow multiple moving bodies, then there is no limit on the degrees of freedom, and hence, the dimension of \( C \). Consider, for example, Figure 17, in which a bunch of rectangles need to be rearranged by translation only. Each contributes 2-D to \( C \). Interestingly, this problem is already NP-hard (and PSPACE-hard) if there is no maximum limit on the number of rectangles. (If the dimension of \( C \) is bounded in advance, then the path-planning problem is solvable in time polynomial in the representation of the robot and world obstacles.)

Planning a collision-free path for multiple rigid bodies is no different conceptually to planning for a single body, once we think in terms of \( C \) and \( C_{\text{free}} \). The configuration vector \( q \in C \) includes coordinates to place each body. For example, for two translation-only rectangles, \( q = (x_1, y_1, x_2, y_2) \) represents their position and \( C = \mathbb{R}^4 \). The initial \( q_i \) and goal \( q_G \) configurations now express the placement of every body. Suppose there are \( n \) bodies \( A_1, A_2, \ldots, A_n \), with configuration parameters \( q_1, \ldots, q_n \). If \( A_i \) is transformed into configuration \( q_i \), it occupies \( A_i(q_i) \subseteq \mathcal{W} \) in the world. Let \( q = (q_1, \ldots, q_n) \) represent the simultaneous configuration of all bodies. A configuration is collision free, \( q \in C_{\text{free}} \), if and only if \( A_i(q_i) \cap \mathcal{O} = \emptyset \) for every \( i \) from 1 to \( n \), and \( A_i(q_i) \cap A_j(q_j) = \emptyset \) for every \( i \neq j \). In other words, for \( q \in C_{\text{free}} \), there must be no body–obstacle collisions and no body–body collisions.

Once \( C, q_i, q_G, \) and \( C_{\text{free}} \) are defined in this way, the methods given in “Combinatorial Planning” and

---

**Figure 16.** The probabilistic road map method attempt to achieve road map accessibility and connectivity preservation via random sampling and connecting to nearby samples.

**Figure 17.** Consider rearranging many rectangles, with no rotations, inside of a rectangular box in \( \mathbb{R}^2 \). Without a limit on the number of rectangles, the problem is NP-hard.
“Sampling-Based Planning” sections directly apply. The only difficulty is that the dimension of \( C \) is large, which limits the applicability of combinatorial methods and some sampling-based methods. This has motivated the development of various decoupled approaches, which avoid considering all bodies at once. For example, paths may be planned for each body individually, and then their motions along the paths can be set correctly so that collisions are avoided. Such methods are not complete but are practical in many settings. Alternatively, dimensionality-reduction techniques, such as those based on the Johnson-Lindenstrauss Lemma, may hold promise for adapting sampling-based planning methods to directly account for all bodies simultaneously.

If bodies are allowed to contact each other, several other motion-planning variants are obtained. Two will be considered here: 1) articulated bodies and 2) manipulation. For articulated bodies, they are attached together by joints that enable some freedom of motion between them, as shown in Figures 18 and 19. The attachment of bodies removes some of their collective degrees of freedom. Configuration coordinates express how each body is situated with respect to bodies to which it is connected. Expressions for transforming such bodies are just standard robot kinematics covered in numerous textbooks [5], [16]. Somewhat different from standard kinematics, we are once again interested in the set of all possible transformations, resulting in the C-space. Once this has been defined, a manifold C-space \( C \) is usually obtained, on which \( q_1, q_2, \) and \( C_{\text{free}} \) are straightforward to define. Here, \( C_{\text{free}} \) includes some configurations in which there are body-body collisions, but only if these they are attached by a joint. Once defined, the methods of “Combinatorial Planning” and “Sampling-Based Planning” sections once again apply, with the usual warning about the dimension of \( C \).

A more serious complication is when a collection of articulated bodies forms a loop, as shown in Figure 20. The result is called a closed kinematic chain, which occurs in parallel robots and if multiple robots contact the same body for manipulation. In most cases, it is difficult to explicitly characterize the set of configurations that satisfy the loop-closure constraint. This makes it difficult to even parameterize paths through \( C \). Sampling-based planning approaches have nevertheless been developed to step through this difficult space by ensuring that loop closure is maintained while incrementally searching for a solution path.

Manipulation problems more generally require robots to determine which bodies to grasp and how to carry them to solve a problem. For example, the task might be to use a manipulator arm to stack several boxes. The degrees of freedom of boxes in addition to the robot are all included when defining \( C \). The task is expressed by specifying a configuration in which the boxes are stacked. This problem conceptually appears more challenging. Standard algorithms are often adapted to solve it by forming a hybrid C-space that includes discrete variables in addition to configuration variables. The discrete variables record modes of interaction. For example, there is a transit mode, when the manipulator is not carrying a body, and a transfer mode, when it carries a body. Heuristics are then used to determine when modes should be switched, in addition to solving the planning problem that arises in each mode.

Another variant of the basic path-planning problem is to allow the obstacles to move. Let \( T = [0, t_f] \) be an interval of time, in which \( t_f \) is some final time. In this case, a snapshot of the world can be imagined at every time \( t \in T \). The obstacle region \( O \) becomes \( O(t) \). Now consider computing a collision-free path from time \( t = 0 \) to time \( t = t_f \). This is

![Figure 18](image1.png) The classic Puma 560 arm is a chain of three rotatable bodies (excluding the end effector) attached to a rigid base. This yields a three-dimensional C-space, which is handled by the standard planning algorithms. (Photo courtesy of the Technical University of Berlin.)

![Figure 19](image2.png) Seven links are attached via rotatable joints. If each is allowed a full range of motion from 0 to \( 2\pi \), then \( C \) is a seven-dimensional torus.

![Figure 20](image3.png) Two or more arms manipulating the same object causes a closed kinematic chain.
conceptually straightforward if we construct the configuration-time space, $Z = C \times T$. Figure 21 shows an example of how this appears. To solve the problem, the path-problem algorithms work in the usual way with one exception: The path must always make forward progress through time. The combinatorial road map methods and incremental sampling and searching methods can be adapted without much difficulty to enforce this. It becomes considerably more challenging, however, if the robot has a maximum speed bound. This yields a constraint on the path slope through $Z$, which is more difficult to enforce. Finally, it is even more difficult and practical, when there is uncertainty in predicting the future motions of the obstacles. This falls under the topic of uncertainty, which is covered in the next tutorial part.

**Conclusions**

After reading this, you should hopefully have extracted the following main points. Motion planning lives in the $C$-space, which is the set of all transformations. Combinatorial planning solves simpler problems in a clean, elegant way, but the running time is too high for industrial-grade problems. Sampling-based planning provides practical solutions for real-world problems but offers weaker guarantees. Performance degrades for problems in which narrow doorways in $C_{\text{free}}$ are hard to find. Several extensions to the standard path-planning problem expand the $C$-space definition and require only minor adaptations to the usual approaches. The key issue is that the $C$-space dimension increases, which generally raises computational complexity.

So we have seen powerful methods that generate a collision-free path automatically. Not bad. This is useful in many settings, extending well beyond robotics. But what if a robot is not able to follow the path due to differential constraints arising from kinematics and dynamics? What if we cannot predict precisely where the robot will go? What if the obstacle locations are uncertain and possibly changing? These concerns, with which every roboticist is familiar, motivate the topics in the second part of this tutorial.

**References**


**Biography**

Steven M. LaValle Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL lavalle@uiuc.edu.
Peccioli, a small medieval town in Italy, became one of the first places in the world where a robot was used (not demonstrated) to carry out a public service in the urban environment (from 15 June 2010 to 7 August 2010). Thirty-five real users accepted to trash their domestic waste using the robot DustCart, a mobile robot designed to collect, transport, and discharge rubbish bags in complete autonomy. During the testing period, the robot safely traveled along the public streets of Peccioli, carrying out its daily service and sharing the urban environment with the passers-by, bicycles, and cars, without causing any problems. Drawing on this unique event, in which the authors also participated, the article addresses some of the implications originating from the actual deployment of autonomous mobile robots in urban areas. Our reflections will gravitate around two major issues: legal regulations and social acceptance. More specifically, we will report on the legal solutions adopted for deploying DustCart in the streets of Peccioli and the activities carried out to increase the social acceptance of the robot.

Till today, the deployment of autonomous mobile robots in urban environments has been the talk of science fiction. A memorable example is a short story and the movie based on it, *I Robot* [1], where the robots carry out various kinds of services in human-inhabited settings. In a particular scene, humanoid robots are walking down the street, shoulder to shoulder with human beings. This is an exemplary case of coexistence between human beings and robots. In this article, we recount a similar story, but this time it is based on real facts: that of a service robot called DustCart, which was used for more than a month in a small Italian town to collect rubbish bags and then transport them to a discharge site. The robot, which was designed and developed within the framework of the European Union (EU) project DustBot [1], traveled on public roads in complete autonomy, interacting with people and cars and coexisting in the urban life of Peccioli. As far as we know, there are no references in literature to service robots being deployed in an urban environment or for such a lengthy period of time.

The objective of this article is to report on the testing of the robot DustCart in Peccioli and to point out some of the ethical, legal, and social implications that emerged before and during the test period.
Ethical Triaging
The different aspects that make autonomous mobile robots such an ethically sensitive topic can be illustrated by drawing on the triage method. The triaging method has been previously used in ethical research in [2] for investigating the ethical implications that arise from the research and deployment of brain–computer interfaces. It has also been used in the framework of the EU-funded project ETHIC-BOTS [3], where it was applied to select ethically sensitive items, namely, technologies or applications that were deemed worthy of ethical investigation. The triaging method consists of analyzing a given technology according to the following two criteria: imminence and social pervasiveness.

As to imminence, it measures the level of maturity and availability of enabling technologies. In our investigation, the enabling technologies refer to the robots’ ability to move autonomously in partially unstructured environments, which include navigation, obstacle avoidance, environmental perception, and self-localization. Recently, these technologies have progressed substantially. A significant illustration of the advancements achieved in the field of autonomous mobile robots is provided by the Defense Advanced Research Project Agency Grand Challenge. The 2007 edition of the challenge included a competition among driverless cars that moved autonomously in an urban environment [4]. Other relevant illustrations of technological maturity in the field of autonomous navigation can be found in a group of projects funded in the sixth framework programme of EU. We refer to the research project DustBot [5], which will be discussed in greater detail in the “Peccioli: The Testing Site for the Robot DustCart” section, ubiquitous networking robotics in urban settings [6], and CyberCars [7]. However, evidence is not only limited to the field of research but also there are a few examples of commercial enterprises that have autonomous vehicles in their product catalogs, such as the French company [8].

The dimension of social pervasiveness deals with the potential impact of a given technology on the society. We argue that autonomous mobile robots may become a very pervasive technology in the near future. Because of their ability to move autonomously in the urban environment, robots can be designed to offer innovative and useful services to human beings. A few general examples of new applications that are still at research level are as follows: guiding people [6], support for the elderly and disabled [10], [11], and a mobile station for monitoring atmospheric pollution [1]. Another interesting field is the provision of solutions to well-known problems that affect urban areas [9], such as reducing road traffic by offering alternative services to people mobility [7], improving rubbish collection and transportation [1], or street cleaning [1]. In private settings, such as factories, there are many types of guide robots that are commercially available [12], [13].

However, a much more reliable indicator of social pervasiveness is given by the growing international market associated with service robotics. Drawing on the figures made available by the International Federation of Robotics in 2010 [14], till 2009, about 77,000 service robots for professional use were sold worldwide and the total value of professional service robots sold was about US$13 billion. In projections for the period 2010–2013, about 80,000 new service robots for professional use will be installed, and the estimated value of sales of service robots for professional use is estimated to be more than US$12 billion. About 30% of the sold units are used for defense applications, 25% for milking, 8% for cleaning, 8% for medical purpose, 7% for underwater robots, 6% each for demolition robots and mobile robot platforms for general use, 5% for logistic systems, and 4% for rescue robots. All these values indicate that robots will become a vital part of our daily life.

Peccioli became one of the first places in the world where a robot was used (not demonstrated) to carry out a public service in the urban environment.

Peccioli: The Testing Site for the Robot DustCart
The testing of DustCart included the implementation of the DustBot system, actually, a slightly modified version of the system developed in the framework of the DustBot project [1]. The DustBot system has already been described in detail [15]. Unlike previous demonstrations of the DustBot system, which lasted only a day and were substantially structured, in Peccioli, for the first time, the system was tested in a real operative environment and with real users. The objectives of the test were to assess the performance of the system, identify its limits, technological as well as those related to the acceptance of the service and the robot by the end users, and to evaluate the economic sustainability of the whole system. The resulting data were necessary to evaluate the feasibility of the industrialization of the DustBot system. In this section, we describe the main elements that made the testing of the robot possible.

The DustCart Robots
The objects of the test were two robot prototypes of the DustCart, which is a mobile autonomous service robot, designed to carry out door-to-door, separate waste collection on demand. The robot consists of a mobile platform, originally the robotic mobility platform 200, a two-wheel robot commercialized by Segway, and customized to support a bin container for the transport and discharge of waste. To make the robot safer and increase its endurance, however, a new mobile base was developed from the scratch for the test in Peccioli. The new base consists of two actuated wheels and two supporting wheels, which overcome a few
technical limitations intrinsic in the Segway platform. The robot was also equipped with two additional powerful batteries that allowed it to work continually for about ten hours. Thanks to the presence of special sensors and other components [16], the robot was able to navigate autonomously avoiding obstacles while moving. With regard to human–robot interfaces, the interaction with human beings took place by means of a touch-screen interface and consisted of simple operations: pressing a graphical button opened and closed the bin container and selecting the corresponding icon on the screen specified the type of waste to be disposed of. The whole interaction procedure was also accompanied by vocal messages. The robot was designed with the usability and acceptance criteria [17] in mind.

**The Service Provided by DustCart**

The service provided by DustCart during the test period in Peccioli was on-demand door-to-door waste collection. The robot was configured to collect three types of waste: undifferentiated, paper, and plastic. The service was in operation from Monday to Sunday, from 8:00 a.m. to 8:00 p.m. except on Tuesdays, when the service was in operation only from 3:00 p.m. to 8:00 p.m. owing to the local market present in the experimental area. To request the DustCart service, users had to call a toll-free number. The calls were managed automatically by the ambient intelligent (AmI) infrastructure (which is described later), which scheduled and allocated the robots. Once a robot was allocated the task, the AmI sent a short message service (SMS) to the user, informing him/her of the arrival time of the robot.

**The Municipality of Peccioli**

Peccioli is a historic village located in the countryside of Tuscany. The village, founded in the Middle Ages, was built on top of a hill and presents the classical topography of a medieval town, with old buildings and narrow, paved streets of various gradients (Figure 1). About 5,000 people live in Peccioli, with a large percentage of elderly people: about 25% of the inhabitants are more than 65 years old. The municipality of Peccioli has a strong penchant for using advanced technologies, with a view to provide its citizens with excellent public services. As a matter of fact, since 1995, the municipality of Peccioli has been collaborating on joint research projects with Scuola Superiore Sant’Anna (SSSA), and as a result, the citizens have had the opportunity to use and test very advanced research and experimentation facilities in the fields of aging, telemedicine, domotics, rehabilitation technologies, energy, environment, wellness, etc.

**The Test Site**

The test site covered three streets and a part of a square. This is the very heart of the town, which is also called ciambellone (which means doughnut) on account of its round shape (Figure 2). The total length of the path selected for the test was approximately 300 m. Within the test site, the streets are almost flat and paved, with shops, bars, restaurants, and other commercial activities on either side. It is worth noting that this is not a pedestrian area, as the roads can still be used for their conventional purpose. The busiest area of the town was selected as the test site to make the robots’ presence strongly noticeable to people, so as to...
better evaluate the robot’s social acceptance and give the word coexistence a concrete meaning.

Within the test site, the control and docking stations were also located. The control station is the place from where human operators supervise the functioning of the DustBot system. In the control station, human operators monitor the robots via remote real-time images that are sent from cameras positioned across the test site (Figure 3). The docking station is the place where the robots stay during the night and where they recharge their batteries and undergo maintenance operations (Figure 4). Near the docking station is the discharge area, which consists of a ramp with a dais, where the robot discharges the collected rubbish. The dais has three holes that correspond to the three typologies of waste collected by the robot (i.e., undifferentiated, paper, and plastic) (Figure 4).

A video surveillance system and a wireless network were installed in the testing site. Wireless coverage was assured by six access points located in the area and connected to the LAN of the control station. The surveillance system consists of six closed-circuit television (CCTV) cameras covering all the experimental areas.
and connected to a central recording system located in the control station.

**The AmI**

In the control station, a PC with a software named *AmI* for managing the system and supervising the activities of the robots was set up (Figure 5).

The users involved in the test were registered in the AmI software, and their telephone numbers were stored in the AmI database along with the robots and collection points. The software allowed us to associate the users to collection points. The AmI software communicates with the robots through a wireless network: when the AmI received a request by phone from a user, it scheduled the first free available robot and tasked the robot with serving the user. While moving, the position of the robot and its status is shown on the map. AmI also allows the operators to stop and resume the motion of the robot in the case of an emergency and/or to cancel the task.

**Recruitment of Test Participants**

The users for DustCart service were recruited from volunteers during a public assembly that was held before the test started and was open to the people of Peccioli. During the assembly, the system and how to use it was explained to the participants and, at the end, 34 users, which consisted of 18 families and 16 commercial activities, agreed to participate in the testing of DustCart. Participants agreed to use only the DustCart robot for the disposal of their rubbish bags for the whole duration of the test period. The majority of the users were 45% retired people and 35% workers. The average number of persons per family is 2.2, and the average age is about 52 years. Before the testing of DustCart, about 50% of the inhabitants of Peccioli separated domestic waste by using traditional bins located in specific areas of the town.

**Legal and Social Reflections**

In addition to technological problems, the major difficulties that had to be overcome to make the testing of DustCart possible were those related to solving legal issues and avoiding social resistance. The following subsections focus on what was done, on the one hand, to solve legal problems related to the robot presence on public roads and, on the other, to improve the robot acceptability among people (Figure 6).

**Legal Regulations**

One of the first questions faced by the organizers of the testing was, “Is it possible to use robots on public roads?” From previous research and studies, we know that there...
exists a legal gap with regard to the juridical status of service robots operating on public roads [18], [19], at the European and presumably international levels. In Europe, according to Article 8 of the “Vienna Convention on Road Traffic,” each moving vehicle, including animals, shall have a driver [20]. In other words, for the road traffic convention, including the Italian highway code, an autonomous vehicle is a contradiction in terms. In what follows, we will discuss the solutions that were adopted and as a result allowed DustCart to operate on the public roads of Peccioli. The municipality in collaboration with the local municipal police took the following measures to ensure safety, avoid traffic congestion, and allow the robot to accomplish its task.

- **Road signs**: Three new road signs were specifically designed for the testing of DustCart.
  
  A) A general warning sign, highlighting the presence of robots operating on the streets. On the sign, there is the following text: “Attention. Area subject to robotic testing.”

  B) A more specific sign, informing the public of the presence of a yellow lane given over to the robot (Figure 7), in which the following text appears: “Attention. Area subject to robotic testing. Yellow lane used by robots.”

  C) Finally, a more specific sign used to warn road users of the possibility of robots crossing the road, which displays the following text: “Attention. Robot crossing. Yellow lane used by robots.”

- **The robot lane**: This lane is a special yellow strip, drawn on the left-hand side of the roads (Figure 8), involved in the test. It was decided that the robots would travel inside the lane, in the direction as that of the cars. The robot lane was meant to separate the robot’s activities from the traffic. It was also meant to prevent cars as well as bikes and bicycles from parking in the robot’s path. To reduce traffic congestion on the narrow streets of Peccioli, three stops were devised on each road to give way to cars.

- **Robot insurance policies**: As the owner of the robots, SSSA was given the task of providing insurance cover for the robot for the duration of the test period. Our first move was to ask our broker whether we would be able to use our insurance policy to cover DustCart against any liability that resulted from the robot’s labor (i.e., death or injury to people or animals). From past experience with public demos, we knew that the SSSA insurance policy covered all research activities, including demos, carried out with our prototypes by the institution personnel, anywhere in the world. However, this time, the broker told us that because of the peculiar nature of the event, the insurance company requested a specific appendix for testing DustCart and the payment of an additional insurance premium. The reason for these requests was the difficulty in placing our robot DustCart within a specific typology transport identified by the Italian highway code.

---

**Figure 6.** A user giving rubbish to the robot. (Photo courtesy of Foto Silvi.)

**Figure 7.** Example of a new road sign (C type), which was specifically designed for the testing of DustCart in Peccioli, Italy. (Photo courtesy of Foto Silvi.)
and this was due to the lack of a human driver. Hence, the robots cannot be subject to the obligatory insurance that covers damage when operating on the road. Eventually, we managed to insure the robot for the test by paying an additional insurance premium of about €800. However, the policy did not cover damages that the robots may sustain. According to our broker, this kind of risk seems to be the most difficult to cover for a company in the insurance market in economically and normative advantageous conditions.

- **Privacy:** Robots, such as DustCart, are endowed with the ability to perceive, store, and use sensitive data related not only to the environment but also to human beings. Therefore, as pointed out by [9], there exists a privacy problem for robots operating in public areas. In Peccioli, the privacy problem was solved by placing in the test site signs warning of the presence of cameras, which were used for security reasons and in accordance with Article 13 of Italian “Personal Data Protection Code” (30 June 2003).

### Social Acceptance

With regard to the presence of an innovation such as DustCart robot in an urban community, our experience shows that social acceptance in this case depended on the following:

1) citizens’ perspective
2) the adoption of a specific innovation process by the local municipality, which was fully supported by our team.

The innovation process—from research to the diffusion and adaptation phases—should take into account the users’ perspectives, in a twofold way. On the one hand, users will adopt (or not) an innovation, awarding its success. On the other hand, they will undergo an innovation, sometimes over their original needs and expectations. Users’ acceptance of innovations such as DustCart, when proposed to an urban community, is also related to two factors: the public’s trust in the local political government and the ability to enable users (citizens) to acquire a balanced view of the new methods required to carry out something they had done differently before. According to the technology acceptance model (TAM) [21], the user’s acceptance of any technology depends on two factors that have a significant impact on a user’s attitude toward using the technology: perceived usefulness and perceived ease of use.

These considerations were taken into account during the two public meetings that were organized by the municipality of Peccioli to involve the citizens in the testing of DustCart. In addition, during the public meetings, the five stages in the decision process identified by [21] were taken into account. As to knowledge and persuasion, all Peccioli citizens were duly informed of the DustCart test. For this purpose, the first public meeting was organized, whose objective was to provide inhabitants of Peccioli with a wide range of information concerning the project and the possible impact of this innovation on Peccioli’s future.

For the same purpose, the organizers carried out a press campaign that included journals and magazines. As a result, some citizens got interested in the test of the robot and actively sought further details about the innovation. For what concerns decision, citizens evaluated the advantages/disadvantages of using the innovation DustCart and decided whether to possibly adopt or definitely reject the innovation. In the testing area, there were 110 users (families and commercial premises): some of them asked to be involved in the test (early adopters [22]), and the first 24, according to the DustBot potentialities, were accepted. An interesting point was the fact that the link with tradition played an important role in accepting the innovation DustCart: people who showed some initial doubts about the robot were definitely convinced as soon as someone remembered them the ancient dustman, named Oscar, who used to collect garbage in a truck when called by Peccioli citizens.
At the implementation stage, usually the individual employs the innovation to a varying degree that depends on the situation and is determined by the usefulness of the innovation and may possibly search for further information about it. In our case, the local municipality organized a second public meeting; this time, the meeting included training on how to use the DustCart robot and request its service and information on its usefulness and ease of use.

Finally, as to confirmation, the citizens finalized their decision to use DustCart for the entire period of the test, using it to its full potential.

By investigating the impact of political factors upon urban policy outputs, the relationship between government and citizens’ decisions has been found to be very important in the adoption of the robot and its service. In this process, the organizational environment and the local government were significant variables and directly influenced the testing of the innovation DustCart, where many other predictors of the use of this innovation would have been expected to be more important. These factors cannot explain or predict the success or failure of the future use of DustCart, but they certainly played a fundamental role in this test.

Conclusions
During the test period, the robot operated for 47 days (from 15 June 2010 to 7 August 2010), for a total of 454 h (with reduced service on Tuesdays), carrying out 382 services and traveling a total of 114.6 km. A total of 560.3 kg of rubbish was collected (paper: 226.5 kg, 40.42%; plastic: 89.7 kg, 16.01%; and undifferentiated garbage: 244.1 kg, 43.57%).

During the testing of DustCart, no accidents occurred, and the robot proved to be reliable and safe.

In this article, we have attempted to highlight that there also exist nontechnological challenges for deploying service robots in urban areas. In particular, we have pointed out the implications related to legal regulations and social acceptance.

As far as the legal issue is concerned, as we have pointed out, the legal classification of autonomous robots at the level of road traffic code still remains a problem. The solutions that were adopted in Peccioli are valid only for a temporary test. With regard to social acceptance, we believe that DustCart was accepted by the inhabitants of Peccioli, and this was on account of the fact that people associated its presence with the accomplishment of an important service: separate waste collection (Figure 9). Citizens also appreciated the fact that DustCart offered an on-demand service; in other words, it fitted the needs of each person and provided a door-to-door service, which meant that people did not have to move away from their homes or shops to dispose of rubbish. This last feature was quite important, especially, for elderly people and for those working in shops.

However, although not a demonstration, but a real deployment, the testing of DustCart was limited in time and took place under partially controlled conditions. In other words, there was neither sufficient time nor the appropriate conditions to find out other potential ethical, legal, and social implications. For instance, it was not possible to find out the existence of abuses or improper behaviors toward the robot, such as vandalism. It has been discussed earlier [23] that vandalism can affect the safety as well as quality of service provided by robots operating in urban environments. A very simple example of robot vandalism is a group of people surrounding the robot and blocking its way. In addition, it would have been interesting to evaluate dustmen’s acceptance level in case the robot was used permanently or that of road users. As discussed earlier, one of the main problems encountered during the testing of the robot was traffic congestion. Despite the presence of a robot lane, the streets of Peccioli are too small for both robots and cars. In conclusion, among the lessons learned from Peccioli that are useful for paving the way for robot deployment in urban settings is that acceptance is no more a matter of the users willingness to use a product [24]. On the contrary, because the robot coexists in a public environment, to determine the overall level of acceptance of innovations, such as DustCart, it should also be necessary to include ethical, social, and legal issues.

Acknowledgments
The DustCart robot was originally developed in the framework of the European Union-funded project DustBot. The authors are grateful to all the people involved in the organization and realization of the DustCart test, especially Belvedere S.P.A. and its president Renzo Macelloni, the mayor of Peccioli Silvano Crecchi.
Peccioli municipal police, Scuola Superiore Sant’Anna, the operators in the control station Silvia Fornai and Marco Toncelli, and RoboTech Srl and Dedalus, the two spin-off companies of SSSA that optimized and adapted the hardware and software components of the DustBot system for the test. Finally, a special thanks to all the citizens of Peccioli who supported and participated in the project.

References


Pericle Salvini, Polo Sant’ Anna Valdera, Viale Rinaldo Piaggio, 34, 56025, Pontedera (PI), Italy. E-mail: p.salvini@sssup.it.

Giancarlo Teti, RoboTech Srl, Scuola Superiore Sant’Anna, Pisa, Italy.

Enza Spadoni, HealthCare and the Biomedical Field at BioRobotics Institute of the Scuola Superiore Sant’Anna, Pisa, Italy.

Cecilia Laschi, Scuola Superiore Sant’Anna (SSSA), Pisa, Italy.

Barbara Mazzolai, Center for Micro-BioRobotics (IIT@SSSA), Istituto Italiano di Tecnologia (IIT), Genoa, Italy.

Paolo Dario, BioRobotics Institute, Scuola Superiore Sant’Anna (SSSA), Pisa, Italy.