CHAPTER

12

Ethical Aspects of Emerging and Converging Technologies

LEARNING OBJECTIVES

Upon completing this chapter, you will successfully be able to:

- Explain what is meant by concept of *technological convergence* and describe how converging technologies raise ethical concerns that can be difficult to anticipate,
- Describe the key components of *ambient intelligence* (*AmI*) and explain why AmI poses some significant social and ethical challenges, especially for personal privacy,
- Describe the key components of *nanotechnology* and explain the social and ethical challenges posed by this relatively recent technology,
- Assess some of the social and ethical impacts that *autonomous machines (AMs)* will likely have in the very near future,
- Understand the key differences between (the relatively new field of) *machine ethics* and (traditional) computer ethics and why that distinction is important,
- Articulate the components of a relatively new "dynamic" ethical framework designed to assess ethical issues that arise in emerging and converging technologies.

In this chapter, the final chapter of *Ethics and Technology*, we examine some ethical and social issues that arise in connection with converging and emerging technologies. We begin by reflecting on a hypothetical scenario that briefly illustrates some ways in which intelligent devices and "smart things" communicate not only with humans but with other devices and things. This phenomenon—as it pertains to one aspect of the emerging and converging technologies examined in this chapter—is now commonly referred to as the *Internet of Things*.

▶ **SCENARIO 12–1:** When "Things" Communicate with One Another

While driving home from work one day, Bill receives a message on his mobile device from the "intelligent refrigerator" in his home, informing him that his supply of milk is very low. This message, in turn, triggers an app (called "Foodster") on Bill's device, which notifies users when grocery items of interest are on sale at selected stores located close to their homes. Foodster informs Bill that milk is on sale at the Sunny Spot convenience store. Next, another app on Bill's device, which communicates with his car's "intelligent driving system," instructs the (auto-enabled) GPS in Bill's car to slightly adjust the route that Bill typically drives home, so he that he will pass by Sunny Spot. When Bill arrives at the store, he notices that gourmet coffee also happens to be on sale that day; so, he queries his intelligent refrigerator about the

amount of coffee currently stored there. The refrigerator then checks Bill's coffee inventory and recommends that he purchase one pound of gourmet coffee while he is shopping at Sunny Spot. But before getting back to Bill with an answer to his original query, the refrigerator first communicates with the kitchen's "intelligent cabinets" to see whether Bill might also need filters for his coffee maker. Bill then receives a message on his mobile device recommending that he purchase three items before leaving Sunny Spot: milk, coffee, and coffee filters. Finally, as Bill approaches the store's checkout line, he clicks on the "shopping rewards" app on his mobile device, which informs him that his CDC Visa card is offering a 5% cash-back option on groceries that he purchases this month; the app then recommends that Bill use his CDC credit card to purchase the items at Sunny Spot. ■

Is this scenario a bit farfetched? Or is it something that would seem realistic in the not-too-distant future? Humans have been interacting with "intelligent agents" and with "smart devices" for several years now. But the idea of intelligent/smart devices and objects communicating with one another, that is, independent of human interaction or human oversight, is a relatively recent phenomenon.

In our discussion of ambient intelligence (AmI) in Section 12.2, we briefly examine some developments involving "smart homes/environments," which many now predict will be commonplace in the near future. There, we will also consider whether these kinds of homes and environments, which may indeed prove to be remarkably convenient in assisting us in carrying out many of our day-to-day tasks, will either improve the overall quality of our lives or have negative consequences for our well-being. The purpose of Scenario 12–1, however, has been simply to get us to begin thinking about what it might be like for many of us in our daily lives, as more and more "intelligent things" inevitably communicate with one another, presumably for the purpose of making our lives easier.

Later in this chapter, we identify and evaluate some social/ethical concerns that arise in connection with nanotechnology and nanocomputing. We also examine some ethical issues affecting "autonomous machines" (AMs) in the context of a relatively new subfield of cyberethics called "machine ethics." In the final section, we describe an ethical framework specifically designed to guide research and inform policies affecting new and emerging/converging technologies. We begin, however, with a brief analysis of the concept of *technological convergence*.

▶ 12.1 CONVERGING TECHNOLOGIES AND TECHNOLOGICAL CONVERGENCE

What, exactly, do we mean by "convergence" in the context of cybertechnology? Howard Rheingold describes technological convergence as a phenomenon that occurs when "apparently unrelated scientific and technological paths" cross or intersect "unexpectedly to create an entirely new field."[1] As we move forward in the twenty-first century, cyber- and non-cybertechnologies are converging at a pace that is unprecedented. However, we saw in Chapter 1 that technological convergence as it pertains to cybertechnology is hardly new. For example, we saw that early computer networks became possible because of the convergence of computing and communication technologies in the late 1960s and early 1970s. Consider that many of the ethical issues we examined in the preceding chapters of this textbook arose because of convergent aspects of computing/information and communication technologies.

Arguably, convergence *within* the domain of cybertechnology itself—that is, the unforeseen blending or merging of disparate, and initially distinct, computing and information technologies (IT)—has been continuous and ongoing. One example of this can be found in virtual reality (VR) technology, which, as Rheingold notes, resulted from the convergence of video technology and computer hardware in the 1980s. (Recall our discussion of technological and ethical aspects of VR in Chapter 11.) However, cyber and cyber-related technologies are now

converging with non-cybertechnologies in ways that challenge our ability to identify and articulate many of the social and ethical issues that also arise either because of or in connection with this kind of convergence.

One ethical/social concern that cuts across the converging technologies examined in this chapter has to do with the new kinds of privacy threats that are now possible. Because Chapter 5 was devoted to privacy concerns pertaining to cybertechnology, you might assume that the appropriate place to discuss these privacy issues would have been in that chapter. There, however, we examined privacy concerns that tend to fit mainly within the category of "informational privacy." For example, those privacy issues typically involve concerns that result from the collection of personal data by commercial and governmental organizations and the mining and analysis of personal information stored in electronic databases. Although some privacy concerns affecting the converging technologies that we examine in this chapter also fall within the category of informational privacy, many do not. The reason for this is not simply because these privacy issues are associated with newer technologies, but because they introduce different kinds of privacy concerns than those examined in Chapter 5.

We will also see that some privacy issues generated by developments in AmI technology and nanotechnologyhave introduced a relatively new category of privacy concern called "location privacy." For example, these converging/emerging technologies can be used to disclose the precise spatial location of an individual at a particular point in time. Because many of the privacy concerns identified and analyzed in this chapter are sufficiently different from those examined in Chapter 5, they warrant a separate context for analysis. However, as you examine the privacy-related issues included in this chapter, you may find it helpful to refer back to relevant sections of Chapter 5. We begin our examination of social and ethical aspects of converging technologies in the twenty-first century with a look at some controversies associated with AmI.

▶ 12.2 AMBIENT INTELLIGENCE (AmI) AND UBIQUITOUS COMPUTING

AmI is often described as a technology that enables people to live and work in environments that respond to them in "intelligent ways."[2] AmI has been made possible, in large part, by the convergence of artificial intelligence (AI) technologies (described in Chapter 11) with (miniaturized) electronic sensing and surveillance technologies. We will examine some technological aspects of AmI in Section 12.2.1; before doing that, however, we may find it useful to recall some concerns briefly described in Scenario 12–1 where "intelligent things" communicated with one another as well as with "Bill's" mobile device. Along somewhat similar lines, Raisinghani et al. (2004) describe a hypothetical case—one that we could call "A Day in the Life of a Smart Home"—where a mother and her child arrive home. As the car pulls into the driveway, the mother is immediately recognized by a surveillance camera that

> disables the alarm, unlocks the front door as she approaches it and turns on the lights to a level of brightness that the home control system has learned she likes.[3]

It turns out that the home control system in this "smart home" has also learned a great deal more about the preferences of its residents. For example, it "knows" when to adjust the room thermostats and how to optimize the use of appliances in order to avoid the risk of power surges occurring at peak hours of electric used in the neighborhood, and so forth.

Is the kind of (smart) home described by Raisinghani et al. based on science fiction? Or does it portray a real-world situation in the not-too-distant future? Consider that a 5,040-square-foot "aware home" was developed at the Georgia Institute of Technology nearly two decades ago; it continues to serve as a laboratory for AmI research and development. Research in AmI has also been conducted at other academic institutions, such as the MIT, as well as at companies in the private sector, such as Philips Electronics. AmI's optimists predict that intelligent homes

will be available to consumers within the next few years. Whereas proponents of AmI are enthusiastic about many of the conveniences made possible by this technology, we will see why critics worry about AmI's "dark side."

We should note that some analysts use the expression "ubiquitous computing," or *ubicomp*, to describe what we refer to in this chapter as AmI. However, ubicomp can easily be confused with "ubiquitous communication," a technological component of AmI. So we use the expression AmI in this chapter to avoid any confusion between the two terms. To better understand AmI technology, we briefly describe three of its key elements or components—pervasive computing, ubiquitous communication, and intelligent user interfaces (IUIs)—before examining some ethical and social aspects of AmI.

12.2.1 Pervasive Computing, Ubiquitous Communication, and Intelligent User Interfaces

According to the Centre for Pervasive Computing (www.pervasive.dk), *pervasive computing* can be viewed as a computing environment where information and communication technology are "everywhere, for everyone, at all times." In this scheme, computing technology is integrated into our environments—from "toys, milk cartons, and desktops to cars, factories, and whole city areas." Pervasive computing is made possible, in part, because of the increasing ease with which circuits can be printed or embedded into objects, including wearable and even disposable items. Pervasive computing goes beyond the traditional scheme of user interfaces—on the one hand, it "implodes them into small devices and appliances"; on the other hand, it "explodes them onto large scale walls, buildings and furniture" (Centre for Pervasive Computing).

Pervasive computing, like AmI, is also sometimes referred to in the computer science literature as *ubiquitous computing* (or *ubicomp*). The expression "ubiquitous computing" was coined by Mark Weiser, who envisioned "omnipresent computers" that serve people in their everyday lives, both at home and at work.[4] He also envisioned ubiquitous computing as something that would function "invisibly and unobtrusively" in the background and that would free people to a considerable degree from tedious routine tasks. For ubiquitous or pervasive computing to operate at its full potential, however, continuous and ubiquitous communication between devices is also needed.

Ubiquitous communication aims at ensuring flexible and omnipresent communication possibilities between interlinked computer devices that can be stationed at various locations. Several different kinds of wireless technologies that make ubiquitous communication possible are now available or are in progress. According to Raisinghani et al. (2004), these include:

- Wireless local area networks (WLANs)
- Wireless personal area networks (WPANs)
- Wireless body area networks (WBANs) interlinking various wearable devices and connecting them to outside networks
- Radio-frequency identification (RFID)

Perhaps the most controversial of these technologies so far—at least from the perspective of personal privacy—is RFID. Koehler and Som (2005) suggest that RFID transponders in the form of "smart labels" will probably become the most widespread example of ubiquitous computing/communication. Recall our discussion of RFID technology in Chapter 5, where we examined some implications of RFID for personal privacy. We will see that RFID technology, when used in AmI environments, can facilitate the tracking of an individual's location at any given point in time and thus make possible a form of "pervasive surveillance."

In addition to pervasive computing and ubiquitous communication technologies, AmI has another key component: *IUIs*. This technology has been made possible by developments in the field of AI. In Chapter 11, we examined AI from the perspective of concerns about our "sense

of self" and about what it means to be a human being in the digital era. There, we also noted that AI, in addition to raising some interesting conceptual and theoretical questions, has many practical applications as well. AI-based applications are also at the core of the "intelligent" user interfaces needed to realize the full potential of AmI.

Brey (2005) notes that IUIs, which are also sometimes called "user adaptive interfaces" because of the way they can adapt to a user's preferences, go beyond traditional interfaces such as a keyboard, mouse, and monitor. As a result, they improve human interaction with technology by making it more intuitive and more efficient than was previously possible with traditional interfaces. With IUIs, for example, computers and electronic devices can "know" and sense far more about a person than was possible with traditional interfaces, including information about that person's situation, context, or environment. Because IUIs respond to inputs such as human gestures as well as to an individual's preferences within various contexts, they enable inhabitants of AmI environments to interact with their environment in a personalized way. Unlike traditional user interfaces, however, IUIs in AmI environments also enable *profiling*, which Brey describes as "the ability to personalize and automatically adapt to a particular user's behavior patterns."

While AmI technology is able to sense changes in an environment and while this technology can automatically adapt and act based on these changes—for example, in response to a user's needs and preferences—AmI remains in the background and is virtually invisible to the user. As Brey notes, people are "surrounded with possibly hundreds of intelligent networked computers that are aware of their presence, personality, and needs." But users themselves may not be aware of the existence of this technology in their environments.

Thus far, we have briefly described three of the key technological components that make AmI possible. We next examine some of the ethical and social challenges posed by AmI environments.

12.2.2 Ethical and Social Aspects of AmI

Social and ethical concerns affecting AmI include worries about the loss of freedom and autonomy. These are sometimes closely related to concerns about humans becoming overly dependent on technology. Other social/ethical concerns involving AmI include threats associated with privacy and surveillance. We begin with a look at some issues affecting freedom and autonomy.

Autonomy, Freedom, and Control

Will human autonomy and freedom be enhanced or diminished as a result of AmI technology? AmI's supporters suggest that humans will gain more control over the environments with which they interact because technology will be more responsive to their needs. However, Brey notes a paradoxical aspect of this claim, pointing out that "greater control" is presumed to be gained through a "delegation of control to machines." But this, he suggests, is tantamount to the notion of "gaining control by giving it away." Brey considers some ways in which control can be gained in one sense and lost in another. With respect to humans gaining control as a result of this technology, he notes that three different kinds of arguments can be made, where AmI may make the human environment more controllable because it can:

1. Become more responsive to the voluntary actions, intentions, and needs of users
2. Supply humans with detailed and personal information about their environment
3. Do what people want without their having to engage in any cognitive or physical effort

On the other hand, Brey considers some ways that AmI can diminish the amount of control that humans have over their environments. These also are organized into three arguments, where users may lose control because a smart object can:

1. Make incorrect inferences about the user, the user's actions, or the situation
2. Require corrective actions on the part of the user
3. Represent the needs and interests of parties other than the user[5]

So, as Brey notes, AmI has the potential to enhance human freedom through its ability to expand certain aspects of our control over the environment—for example, in responding to our voluntary actions, intentions, and needs and by freeing us from many routine and tedious tasks that require either cognitive or physical effort. But he also notes that AmI has the potential to limit freedom because it can make incorrect inferences about a user's intentions and needs. Even when AmI does what a user wants, it can still reduce control by requiring "corrective actions" on the part of the user. Brey also notes that users can lose control when smart objects perform autonomous actions that do not solely represent the user's interests. For example, the smart object could include a user profile or knowledge base that is also designed to take into account the interests of third parties (such as commercial interests). Additionally, Brey believes that AmI could undermine human freedom and autonomy if humans become too dependent on machines for their judgments and decisions.

Technological Dependency

We have come to depend a great deal on technology, especially on digital technology, in conducting many activities in our day-to-day lives. In the future, will humans depend on the kind of smart objects and smart environments made possible by AmI technology in ways that exceed our current dependency on computing and electronic devices? We noted earlier that IUIs could relieve us of having to worry about performing many of our routine day-to-day tasks, which can be considered tedious and boring. But we also noted that these interfaces could relieve us of much of the cognitive effort that has, in the past, enabled us to be fulfilled and to flourish as humans. What would happen to us if we were to lose this capacity because of an increased dependency on technology? Perhaps a brief look at a scenario envisioned by E. M. Forster in one of his classic works would be instructive at this point.

▶ **SCENARIO 12–2:** E. M. Forster's "(Pre)Cautionary Tale"

In his short story *The Machine Stops*, first published in 1909, E. M. Forster portrays a futuristic society that, initially at least, might seem like an ideal or utopian world. In fact, Forster's story anticipated many yet-to-be-developed technologies such as television and videoconferencing. But it also illustrates how humans have transferred control of much of their lives to a global Machine, which is capable of satisfying their physical and spiritual needs and desires. In surrendering so much control to the Machine, however, people begin to lose touch with the natural world. After a while, defects appear in the Machine, and eventually it breaks down. Unfortunately, no one remembers how to repair it. In Forster's tale, some of the characters begin to realize just how dependent they have become on this machine.[6] ■

We could easily imagine Forster's scenario playing out in AmI environments of the future where individuals no longer are required to perform routine cognitive acts and instead depend on IUIs to make decisions for them. Also, we could ask what would happen if the energy sources that powered the AmI environments were suddenly lost. Could we respond successfully if this happened to us? If not, it would seem that we have let ourselves become too dependent on this technology. Hypothetical questions of this kind are worth keeping in mind as we proceed with developments in AmI.

Privacy, Surveillance, and "the Panopticon"

Some of AmI's critics worry that a kind of "Big Brother" society may emerge. For example, Bohn et al. (2005) note that in AmI environments, all of our moves, actions, and decisions will be recorded by "tireless electronic devices, from the kitchen and living room of our homes to

our weekend trips in cars." As Langheinrich (2001) points out, no aspect of our life will be secluded from "digitization," because virtually anything we say, do, or even feel could be "digitized, stored, and retrieved anytime later." But how are the privacy concerns associated with AmI different, in relevant ways, from privacy issues generated by earlier uses of computer/information technology? Langheinrich believes that with respect to privacy and surveillance, four features differentiate AmI from other kinds of computing/IT applications:

- Ubiquity
- Invisibility
- Sensing
- Memory application

First, Langheinrich believes that because computing devices are *ubiquitous* or omnipresent in AmI environments, privacy threats involving AmI are more pervasive in scope and can affect us more deeply. Second, because computers are virtually *invisible* in AmI environments (in the sense that they easily "disappear" from view), it is likely that users will not always realize that computing/electronic devices are present and are being used to collect and disseminate personal data. Third, *sensing* devices associated with the IUIs in AmI environments may become so sophisticated that, unlike conventional forms of cybertechnology, they will be able to sense (private and intimate) human emotions such as fear, stress, and excitement. Fourth, AmI has the potential to create a *memory* or "life-log"—that is, a complete record of someone's past. So, Langheinrich concludes that AmI poses a more significant threat to privacy than earlier computing/information technologies.

In AmI environments, the sheer scale or amount of information that can be collected without our awareness is also problematic. Bohn et al. note that AmI has the potential to create a comprehensive surveillance network, because it can disclose an "unprecedented share of our public and private life." We saw that AmI environments are equipped with sensors that facilitate the collection of data about an individual from his or her surroundings without that individual's active intervention. This kind of ubiquitous observation, which some now call "pervasive surveillance," can expose much about an individual's habits and preferences.

Čas (2005) notes that no one can be sure that his or her actions are not being observed; nor can one be sure that his or her words are not being recorded. Furthermore, individuals cannot be sure whether information about their presence at any location is being recorded. So, he believes that the only realistic attitude of human beings living in such environments is to assume that any activity or inactivity is being monitored, analyzed, transferred, and stored and that this information may be used in any context in the future. In this sense, people in AmI environments would be subject to a virtual "panopticon."

▶ SCENARIO 12–3: Jeremy Bentham's "Panopticon/Inspection House" (Thought Experiment)

Jeremy Bentham, an eighteenth-century philosopher and social reformer, conceived of the idea for managing a prison environment based on the notion of the *panopticon*. Imagine a prison comprised of glass cells, all arranged in a circle, where prisoners could be observed at any moment by a prison guard who sits at a rotating desk facing the prisoner's cells. Further imagine that the inmates cannot see anyone or anything outside their cells, even though they can be observed (through the one-way-vision glass cells) by the prison guard at any time. Although a prisoner cannot be certain that he is being observed at any given moment, it would be prudent for him to assume that he is being observed at every moment. The prisoner's realization that he could be observed continuously, and his fear about what could happen to him if he is observed doing something that is not permitted in the cell, would likely be sufficient to control the prisoner's behavior.[7] ■

TABLE 12-1 Ambient Intelligence

Technological Components	Ethical and Social Issues Generated
Pervasive computing	Freedom and autonomy
Ubiquitous communication	Privacy and surveillance
Intelligent user interfaces	Technological dependence

Suppose Bentham's model of the panopticon or "inspection house" were to be extended to public spaces including public buildings. Further suppose that it is extended to include private and intimate environments as well. What effects could the possibility of being permanently observed have on individual behavior and social control? In Bentham's classical panopticon, one could not be certain whether he or she was actually being monitored at a given point in time. Persons living in AmI environments, however, can, with almost 100% certainty, know that they are being observed. Classical forms of surveillance, from Bentham's time to the period preceding AmI technology, were limited to time and place. But data captured in AmI environments will, as Čas notes, persist across space and time.

So far, we have examined a cluster of social and ethical concerns affecting AmI environments. Table 12-1 lists the technological components of AmI and the corresponding ethical and social issues associated with them.

We have seen that some of these ethical and social issues arise because of the pervasive aspects of AmI technology, while others reflect concerns pertaining to convergent features of its component technologies. In the next section, we examine some ethical concerns that result from converging aspects of computing and nano technologies—that is, controversies at the intersection of cybertechnology and nanotechnology. Chadwick and Marturano (2006) argue that nanotechnology provides the "key" to technological convergence in the twenty-first century.

▶ 12.3 NANOTECHNOLOGY AND NANOCOMPUTING

What, exactly, is nanotechnology? Why is research at the nanolevel controversial from an ethical perspective? Should we continue to engage in research and development in nanocomputing? We examine each of these questions, beginning with an overview of nanotechnology as a scientific field.

12.3.1 Nanotechnology: A Brief Overview

Berne (2015) describes *nanotechnology* as "the study, design, and manipulation of natural phenomena, artificial phenomena, and technological phenomena at the nanometer level." We should note, however, that, at this time, there is no universally agreed-upon definition of the field. One common or unifying feature of nanotechnology, regardless of how narrowly or broadly it is defined, is that it operates on matter on a scale of *nanometers* (nm).

Moor and Weckert (2004) note that a nanometer, which is one billionth of a meter, is very close to the dimensions of individual atoms whose diameters range from 0.1 to 0.5 nm. K. Eric Drexler, who coined the term "nanotechnology" in the 1980s, conceived of the field as a branch of engineering dedicated to the development of electronic circuits and mechanical devices built at the molecular level of matter (Drexler 1986). Although such nanolevel devices do not yet exist, current microelectromechanical systems (MEMS), tiny devices such as sensors embedded in conductor chips used in airbag systems to detect collisions, are one step away from the molecular machines envisioned by Drexler.

The Development of Nanotechnology as a Field of Scientific Research

The origin of nanotechnology as a distinct field is generally traced to a 1959 talk by physicist and Nobel laureate Richard Feynman, who encouraged scientists to develop tools that could manipulate matter at the atomic level. In 1990, Donald Eigler and Erhard Schweizer, two scientists working at the IBM Almaden laboratory, succeeded in manipulating 35 individual xenon atoms to shape the three initials of their employer's logo. Since then, more practical kinds of applications have been carried out at the nanolevel. Drexler has proposed the idea of a nanoscale *assembler*—that is, a molecular machine that could be programmed to build virtually any molecular structure or device from simpler chemical building blocks. He believes that the development of universally applicable assemblers, which could be programmed to replicate themselves, is essential for the full realization of nanotechnology's potential.

Although some critics argue that nanotechnology has generated more hype than substance, a few important breakthroughs have already begun to occur at the nanolevel. For example, Regis (2009) describes some of the implications of the nanotube radio that was invented by Alex Zettl and his colleagues in 2007. Regis notes that a "single carbon nanotube tunes in a broadcast signal, amplifies it, converts it to an audio signal and then sends it to an external speaker in a form that the human ear can readily hear." He also notes that this could be the "basis for a new range of applications: hearing aids, cell phones, and iPods small enough to fit completely within the ear canal."[8]

Nanocomputers and Nanocomputing

In the 1980s, Drexler predicted that developments in nanotechnology would result in computers at the nanoscale—that is, *nanocomputers*. Merkle (1997) believes that future nanocomputers will have mass storage devices capable of storing more than 100 billion bytes in a volume the size of a sugar cube and that these devices will be able to "deliver a billion billion instructions per second." Drexler (1991) suggests that nanocomputers will be designed using various types of architectures. For example, an electronic nanocomputer would operate in a manner similar to present-day computers, differing primarily in terms of size and scale. A quantum nanocomputer, on the contrary, would work by storing data in the form of atomic quantum states or spin. Weckert (2006) notes that quantum computers would be much more powerful than any computing systems available today.

Some predict that future nanocomputers will also be built from biological material such as DNA. For example, Seeman (2004) believes that DNA is an ideal molecule for building nanometer-scale structures because strands of DNA can be "programmed to self assemble into complex arrangements" that bond together. And Drexler, who believes that biology shows us how molecular machinery can construct complex organisms from the bottom up, suggests that biological computers are already a reality.

Whether biological and quantum computers will be functionally available at the nanolevel is still a matter of conjecture and debate. However, more conventional notions of computing at the nanoscale are currently under development, and some standard computing chips have already been constructed at the nanoscale. At Hewlett Packard, for example, researchers have made computer memory devices by creating eight platinum wires 40 nm wide on a silicon wafer. Moor and Weckert note that it would take more than 1,000 of these chips to be the width of a human hair.

Before identifying and analyzing the ethical aspects of nanocomputing and nanotechnology, a principal objective of Section 12.2, we should point out that nanotechnology's optimists and pessimists have been quick to offer their predictions about the societal advantages and disadvantages that could result from continued nanotechnology development. For example, Gordijn (2003) notes that optimists point to some of the advantages for the medical field (with nanobots assisting in surgery), while pessimists describe some "apocalyptic nightmares" that could result (including nanolevel weapons and destruction). Weckert (2006) believes that

because many predictions about nanotechnology seem reasonable, it would be prudent for us to consider some of the ethical implications now while there is still time to anticipate them.

12.3.2 Ethical Issues in Nanotechnology and Nanocomputing

Moor and Weckert (2004) believe that assessing ethical issues that arise at the nanoscale is important because of the kinds of "policy vacuums" (Moor 2001) that can arise. (Recall our discussion of Moor's notion of policy vacuums in Chapter 1.) Although Moor and Weckert do not explicitly argue that a separate field of applied ethics called *nanoethics* is necessary, they make a convincing case for why an analysis of ethical issues at the nanolevel is now critical. In particular, they identify three distinct kinds of ethical concerns that warrant analysis:

1. Privacy and control
2. Longevity
3. Runaway nanobots

With respect to (1), the authors note that as we construct nanoscale information gathering systems, it will become extremely easy to put a nanoscale transmitter in a room or onto someone's clothing in such a way that he or she will have no idea that the device is present or that he or she is being monitored and tracked. Implanting tracking mechanisms within someone's body would also become easier with nanotech devices. Moor and Weckert note that a tracking mechanism might be put into someone's food so that, when swallowed, it would be absorbed into the body, possibly migrating to a desired location. The authors further note that in addition to privacy threats made possible by nanotechnology, individuals may also lose some degree of control. Because other people could know more about each other, for example, we might be less capable of controlling the outcomes of our choices. How these tracking devices will be developed and used is still a matter of some speculation. But Moor and Weckert argue that with the advent of nanotechnology, invasions of privacy and unjustified control over others will most likely increase.

Regarding (2), ethical concerns involving longevity, Moor and Weckert argue that developments in nanotechnology could have a dramatic effect on human life spans. While many see longevity as a good thing, there could be negative consequences as well. For one thing, Moor and Weckert note that there could be a population problem if the life expectancy of individuals were to change dramatically. The authors also point out that if fewer children are born relative to adults, there could be a concern about the lack of new ideas and "new blood." Additionally, questions could arise with regard to how many "family sets" couples, whose lives could be extended significantly, would be allowed to have during their expanded lifetime. Other questions might be conceptually confusing—for example, would the (already) old stay older longer, and would the young remain young longer? So, in Moor and Weckert's analysis, longevity-related questions introduce some policy vacuums, as well as conceptual muddles, that will need to be resolved.

With regard to (3), Moor and Weckert argue that we need to consider the potential problem of "runaway nanobots." (The problem of runaway replication in the context of nanotechnology is often referred to as the "grey-goo scenario.") Moor and Weckert note that the replication of these bots could get out of hand. The authors also note that when nanobots work to our benefit, they build what we desire. But when they work incorrectly, they build what we don't want.

Some critics, including Smalley (2001), have challenged the possibility of replicators, because of the way these assemblers would have to be constructed. Drexler, however, responds to Smalley's challenges by noting that biological assemblers such as ribosomes already do the

kind of assembly at the molecular level needed for nanobots. Woodhouse (2004) notes that important choices about how to proceed with nanotechnology will have to be made before it is determined whose prediction—Drexler's or Smalley's—is correct. So, as long as it may be possible to construct nanolevel robots that are capable of self-assembly and replication, it would be prudent to try to anticipate the ethical outcomes that could arise.

Should Nano Research/Development Continue?

While we have examined some ethical concerns associated with potential developments at the nanolevel, we have not yet directly addressed the implications that these developments can have for computer scientists and computing/IT professionals working on nanolevel projects. In Chapter 4, we examined some ethical challenges that computing/IT professionals face. However, we did not discuss any nanocomputing-specific issues there. Next, we identify some of those challenges.

We begin by noting that Joseph Weizenbaum (1976) argued that there are certain kinds of computer science research that should not be undertaken—specifically research that can easily be seen to have "irreversible and not entirely unforeseeable side effects." Weizenbaum did not refer to nanotechnology research per se; however, Joy (2000), who has since echoed some of Weizenbaum's concerns about technological research, worries that because developments in nanocomputing threaten to make us an "endangered species," the only realistic alternative is to limit the development of that technology. Others, however, such as Merkle (2001) disagree with Joy. Merkle argues that if research in nanocomputing and nanotechnology is prohibited, or even restricted, it will be done underground. If that happens, Merkle worries that nanotechnology research would not be regulated by governments and professional agencies concerned with social responsibility.

If Joy and others are correct about the dangers of nanotechnology, we must seriously consider whether research in this area should be limited and whether computer scientists should participate in developments in nanocomputing. However, major computing associations such as the ACM and IEEE have not taken a stance on questions involving the ethics of nanocomputing research and development. Should research in this area be sanctioned by professional computing associations? If not, should nanocomputing research continue? What kind of criteria should be used in establishing a coherent nanotechnology policy?

Initially, we might assume that because nanotechnology could be abused—for example, used to invade privacy, produce weapons, etc.—nanocomputers should not be developed, or at least their development should not be sanctioned by professional computing/IT associations. However, we would commit a logical fallacy (see the Slippery Slope Fallacy in Chapter 3) if we used the following kind of reasoning: Because some technology, X, could be abused or because using Technology X could result in unintended tragedies, X should not be allowed to be developed. Consider some examples of why this form of reasoning is fallacious. Automobiles and medical drugs can both be abused, and each can contribute to the number of unintended deaths in a given year, even when used appropriately. In the United States, more than 40,000 deaths result each year from automobile accidents. And medical drugs (designed to save lives) have also been abused by some individuals, which has resulted in many deaths each year. Should the development of automobiles have been banned? Should we stop research on medical drugs? It would be fallacious to conclude that we should ban the development of these products merely because they could be abused and because they will inevitably lead to unintended deaths.

Arguments for how best to proceed in scientific research when there are concerns about harm to the public good, especially harms affecting the environmental and health areas, are often framed and evaluated via a scheme known as the "precautionary principle." We next examine that principle in the context of nanotechnology.

Assessing Nanotechnology Risks: Applying the Precautionary Principle?

Clarke (2005) notes that many formulations of the *precautionary principle* have been used in the scientific community; so there is no (single) universally agreed-upon formulation of this important principle. According to Weckert and Moor (2004), however, the essence of the precautionary principle can be captured and expressed in the following way:

> If some action has a possibility of causing harm, [it] should not be undertaken or some measure should be put in its place to minimize or eliminate the potential harms.[9]

Weckert and Moor believe that when the precautionary principle is applied to questions about nanotechnology research and development, it needs to be analyzed in terms of three different categories of harm: "direct harm," "harm by misuse," and "harm by mistake or accident." With respect to direct harm, they analyze a scenario in which the use of nanoparticles in products could be damaging to the health of some people. Weckert and Moor note that the kinds of risks in this scenario are very different from those used in the example they select to illustrate harm by misuse—namely, that developments in nanoelectronics could endanger personal privacy. Here, it is neither the new technology nor the product itself that could cause the problem, but rather the way that the new technology/product is used. Weckert and Moor also note that in this scenario, preventing certain uses of the technology would avoid the problem, without stopping the development of nanotechnology itself.

Regarding the third category, harm by mistake or accident, Weckert and Moor describe a scenario in which nanotechnology could lead to the development of self-replicating, and thus "runaway," nanobots. The authors note that harm will occur in this scenario *only if* mistakes are made or accidents occur. But this kind of potential harm is very different from the kind that results from the development of products that will damage health or from technologies that can be deliberately misused. Whereas legislation can be enacted to stop inappropriate uses of a technology or to prevent the development of products known in advance to be harmful to one's health, it is more difficult to draft legislation that will control mistakes and accidents.

Weckert and Moor conclude that when assessing the risks of nanotechnology via the precautionary principle, we need to look at not only potential harms and benefits of nanotechnology per se but also at the "relationship between the initial action and the potential harm." In their scenario involving direct harm, for example, nanoparticles damaging health, the relationship is fairly clear and straightforward: We simply need to know more about the scientific evidence for nanoparticles causing harm. But in their scenario involving potential misuse of nanotechnology, for example, in endangering personal privacy, the relationship is less clear. Here, we need scientific evidence that certain kinds of devices can be developed, and we need evidence about whether effective legislation could be implemented to control the uses of the devices. In their third scenario, we need evidence regarding the propensity of humans to make mistakes or the propensity of accidents to happen.

So, given the risks and potential harms that could result from future developments in nanotechnology, how should research in that field proceed? Weckert (2006) believes that, all things being equal, *potential* disadvantages that can result from research in a particular field are not in themselves sufficient grounds for halting research altogether. Rather, he suggests that there should be a "presumption in favor of freedom in research" until it can be clearly shown that the research is, in fact, dangerous. However, once a reasonable (or what he calls a "*prima facie*") case can be made to show that the research is dangerous, the burden for showing that the research is safe (and that it should continue) would shift from those who oppose the research to those who support it. In Weckert's view, then, it would be permissible to restrict or even forbid research in a field where it can be clearly shown that significant harm is more likely than not to result from that research.[10]

Using Weckert's model, it would seem that since there are no compelling grounds (at present) for halting nanotechnology and nanocomputing research, we should proceed with it. Of course, we would need to reassess our default presumption in favor of nanotechnology/nanocomputing research, if evidence in the future were to suggest that such research posed a serious threat to our safety. We elaborate on this important point in Section 12.6, where we examine a "dynamic" model of ethics that takes into account the need to update factual data as it becomes available, as part of the ongoing process of ethical evaluation. Next, however, we consider some ethical aspects of a different kind of emerging technology: autonomous machines (AMs).

▶ 12.4 AUTONOMOUS MACHINES

Thus far, we examined ethical aspects of two relatively recent technologies that have emerged as a result of converging technological components: AmI and nanocomputing. In this section, we consider an emerging technology that has been made possible, in large part, by recent developments in AI and robotics—namely, *AMs*. We begin our analysis by defining some key terms, as well as drawing some important conceptual distinctions, regarding the various technologies and systems associated with AMs.

12.4.1 What is an AM?

For our purposes, an *AM* is any computerized system/agent/robot that is capable of acting and making decisions independently of human oversight. An AM can also (i) interact with and adapt to (changes in) its environment and (ii) learn (as it functions).[11] We use the expression "autonomous machine" in a broad sense to include three conceptually distinct, but sometimes overlapping, autonomous technologies: artificial agents (AAs), autonomous systems, and robots. The key attribute that links or brings together these otherwise distinct (software) programs, systems, and entities is their ability to act *autonomously*, or at least act independently of human intervention.

Autonomous Machines vs. Autonomous Robots/Agents/Systems

Why use "AMs" rather than "robots," "autonomous artificial agents," or "autonomous systems" to describe the autonomous technologies described in this section? For our purposes, there are two reasons why the phrase "autonomous machine" is more appropriate than "robot." First, not all robots are autonomous, and thus capable of acting independently of humans. Sullins (2011) distinguishes between "tele robots," which are controlled remotely by humans (and function mainly as tools), and "autonomous robots" which can make "major decisions about their actions using their own program." Second, the term "robot" can be ambiguous, because "soft" bots (such as AI programs) are also sometimes included under the general category of robot. To avoid this ambiguity, Wallach and Allen use the expression "(ro)bot." However, our notion of "autonomous machine" is sufficiently robust to capture both the breadth of Wallach and Allen's "(ro)bot" and the precision needed to exclude Sullin's category of (non-autonomous) telerobots.

The expression "autonomous machine" also has an advantage over the phrase "autonomous artificial agent." For one thing, "machine" can be a less philosophically controversial category than "agent" or "artificial agent" (AA); for another, "machine" is a sufficiently broad category to subsume under it certain kinds of entities, systems, etc. that may not fall neatly into the categories of agent and AA. Also, distinctions between a single AA and multiple AAs, such

as "multi-agent systems," can be problematic from the philosophical perspective of agency. However, our category of "autonomous machines" can be understood to subsume both individual AAs and collections of AAs, including multiagent systems.

Third, and finally, "autonomous machine" also has an important advantage over "autonomous system." One problem with the latter expression is that it is ambiguous and can easily be used equivocally to refer to two very different kinds of technologies. On the one hand, an autonomous system (AS), in the context of the Internet, refers to a collection of Internet protocol (IP) routers or "routing prefixes" that are "under the control of one or more network operators"—in this case, an AS can be either a network or set of networks that is "controlled by a common network administrator."[12] On the other hand, "autonomous system" is also used to describe a computerized system that, like an AM, can operate without human intervention, adapt to its environment, learn (as it functions), and make decisions.[13] So, we use the expression "autonomous machine" to avoid the potential equivocation that can easily arise in discussions involving ASs, given the two common uses of "autonomous system." For our purposes, the phrase "autonomous machines" both (i) captures the second sense of "autonomous system," as described in the Royal Academy of Engineering's 2009 report, and (ii) eliminates any ambiguity or equivocation that can arise because of the first sense of AS (i.e., in connection with Internet router policies).

Understanding What Is Meant by "Machine"

Of course, it is possible that some might object to our use of "machine" because that concept usually connotes something physical, as in the case of computer hardware. In this sense, "machine" might be interpreted in a way that would exclude software (programs and applications). So a more precise, and perhaps also more expanded, definition of what is meant by a *machine* is needed in the case of our category of AMs. Even though we tend to think of machines primarily as physical devices consisting of fixed and movable parts, a machine can also be understood as a "natural system or organism." It can also refer to a group of individuals that are under the control of a leader, such as in the case of a "political machine."[14] So, "machine" can be used in both a physical and a non-physical sense. While robots clearly fit within the former sense of "machine," the term's latter sense can include AI (soft)bots, AAs, and ASs that are non-physical. Thus, an AM, as we use the phrase, includes both senses of "machine."

Hall (2011) argues that the most important "machine" of the twentieth century was not a physical entity at all; rather, it was a "Turing Machine," which he describes as a "theoretical concept of a pattern of operations that could be implemented in a number of ways." Hall also notes that a Turing machine can be viewed as a "mathematical idea" that provided the "theoretical basis for a computer." It can also be viewed as a kind of "virtual machine"; in this scheme, any program running on a computer is also a virtual machine. But Hall believes that we can eliminate the "virtual" in these kinds of machines and refer to computer programs themselves simply as "machines." He argues that the essence of a machine is "its behavior"—that is, "what it does given what it senses."[15] In this sense, AMs can also be viewed as machines (and not merely as virtual machines).

Finally, we should note that because AMs have been made possible by developments in AI, "intelligence" is an essential feature or property of AMs. In fact, this feature can also help us to distinguish AMs from what we might think of as ordinary or conventional machines, including some physical devices that are fairly sophisticated. However, it is also important to note that not every "intelligent machine" is necessarily autonomous. We examine some key criteria that (intelligent) machines must satisfy to act "autonomously" in our analysis of the concept of autonomy in Section 12.5.2. First, however, we identify some typical examples of AMs.

Some Examples and Applications of AMs

A highly influential report (on autonomous systems) by the UK's Royal Academy of Engineering (2009) identifies various kinds of devices, entities, and systems that also fit nicely under our category of AM. These include:

- Driverless transport systems (in commerce)
- Unmanned vehicles in military/defense applications (e.g., "drones")
- Robots on the battlefield
- Autonomous robotic surgery devices
- Personal care support systems

Another example identified in that report is a "smart environment," such as a "smart" building/home/apartment. (Recall the example of a hypothetical "smart home" that we briefly described in Section 12.2 in our discussion of AmI; that technology also qualifies as a kind of AM.) Other examples of AMs include driverless trains that shuttle passengers between terminals in large airports, as well as robotic companions/caregivers that assist the elderly and robotic babysitters (which are popular in Japan) that entertain young children.

A diverse cluster of AMs now function in multiple sectors of our society. Consider, for example, the many different kinds of robots and robotic systems that have become available in recent years. Lin (2012) identifies a range of sectors in which robots (and, in our case, AMs) now operate; these include:

1. Labor and service
2. Military and security
3. Research and education
4. Entertainment
5. Medical and healthcare
6. Personal care and companionship[16]

Lin points out that an example of an AM used for (1) would be the Roomba vacuum cleaner, and he notes that nearly half of the 7-million-plus service robots in the world are Roombas. We should point out that while Roombas may appear to act autonomously because of their sensing abilities, they are still also under human control. However, the Roomba, which is probably better viewed as a kind of semi-AM, can still be viewed as a major advancement over earlier industrial robots that operated in automobile factories and assembly lines.

Examples of AMs used in (2) would include the U.S. military's Predator and BigDog, whereas an instance of an AM used in (3) is NASA's Mars Exploration Rover. Lin identifies ASIMO (Advanced Step in Innovative Mobility), a humanoid robot designed by Honda, as an example of an AM that can be used in (4), and he describes some robotic nurses (including RIBA) and robotic pharmacists (such as ERNIE) as examples of AMs used in (5). Lin notes that AMs used in (6) would include CareBot and PALRO, and he also notes that this category of robots might be extended to include some recently introduced "sex bots" such as Roxxxy.

Despite the many conveniences and services that AMs provide, these machines raise some ethical concerns (as we have already noted). One such concern involves threats to personal privacy. Consider that some kinds of AMs allow for detailed recording of personal information; for example, people who live in "smart apartments" could have vast amounts of personal information about them recorded and kept by a third party. The privacy concerns that arise here are very similar to the kinds of AmI-centered privacy issues we examined in Section 12.2.2. Because AM-related privacy concerns overlap with those involving AmI, we

will not examine any AM-specific privacy issues in the following section. Instead, we will focus on three very different kinds of ethical/philosophical concerns affecting AMs: (moral) agency, autonomy, and trust.

12.4.2 Some Ethical and Philosophical Questions Pertaining to AMs

Some ethical issues associated with AMs also cut across traditional cyberethics categories such as property, privacy, security, and so forth. For example, we have already noted that privacy concerns can arise in connection with specific kinds of AMs (such as "smart homes"). Another cluster of ethical concerns involve moral and professional responsibility issues associated with designing AMs. We briefly examine some of those concerns in Section 12.5.2. However, some questions that arise in connection with AMs are not only ethical in nature but are also more broadly philosophical (e.g., metaphysical or epistemological). These include questions about agency (and moral agency), in connection with concerns about whether AMs can be held responsible and blameworthy in some sense, as well as questions about autonomy and trust.[17] We begin by asking in which sense(s) an AM can be viewed as an agent, or artificial agent, before considering the more controversial question of whether an AM can qualify as a moral agent.

AMs, Agents, and Moral Agents?

As already noted, the concepts of "agency" and "agent" can be philosophically controversial. For our purposes, however, we can stipulate a definition of *agent* as someone or something that is capable of acting. So, each of us, insofar as we can act, qualifies as an agent; other entities—both humans and non-humans—who act on our behalf also qualify as agents (and are sometimes referred to as "fiduciary agents"). We refer to all non-human agents as AAs. In our scheme, even a thermostat can satisfy the conditions for being an AA. Today, AI researchers typically refer to artificial entities—whether software programs (in the form of "bots") or full-fledged robots—as AAs.

Because AMs are capable of acting, they also qualify as AAs. But unlike low-level AAs such as thermostats, AMs can act in ways that have a moral impact. So it might seem reasonable to ask whether we can hold AMs morally accountable for their actions. Initially, this might seem like a bizarre question. However, one concern raised in the Royal Academy's influential report on autonomous systems (2009) is whether systems like AMs should be regarded as "robotic people," as opposed to mere machines. This question is important because if AMs qualify as "people" of some sort, they could also be subject to (moral) blame for faults that occur, as well as for legal liability in cases involving either the deaths of humans or severe economic losses. Although it might seem odd to talk about AMs as "people," robotic or otherwise, we have seen that they do qualify as agents—namely, AAs. But can AMs also satisfy the additional conditions that are required for being *moral agents*?

Floridi (2011) believes that AMs, or what he calls autonomous AAs, can be moral agents because they are (i) "sources of moral action" and (ii) can cause moral harm or moral good. In Chapter 11, we saw that Floridi distinguished between "moral patients" (as receivers of moral action) and moral agents (as sources of moral action). There, we also noted that information entities, in Floridi's view, deserved consideration (minimally) as moral patients, even if they were not moral agents. But, additionally, Floridi believes that autonomous AAs also qualify as moral agents because of their (moral) efficacy. Johnson (2006) also believes that AAs have moral efficacy, but she argues that they qualify only as "moral entities" and not moral agents because AAs lack freedom. And others, including Himma (2009), argue that because these entities also lack consciousness and intentionality, they cannot satisfy the conditions for moral agency.

Moor (2006) takes a different tack in analyzing this controversial question by focusing on various kinds of "moral impacts" that AAs can have. Moor begins by noting that computers

can be viewed as normative (non-moral) agents, independent of whether they are also moral agents, because of the normative impacts their actions have. He points out that computers are designed for specific purposes and thus can be evaluated in terms of how good or how bad they perform in accomplishing the tasks they are programmed to carry out (e.g., as in the case of a program designed to play chess). Moor then notes that some normative impacts made possible by computers can also be moral or ethical in nature, and he argues that the consequences, and potential consequences, of what he calls "ethical agents" can be analyzed in terms of four levels:

- Ethical impact agents
- Implicit ethical agents
- Explicit ethical agents
- Full ethical agents

Moor notes that whereas ethical impact agents (i.e., the weakest sense of moral agent) will have ethical consequences to their acts, implicit ethical agents have some ethical considerations built into their design and "will employ some automatic ethical actions for fixed situations." And while explicit ethical agents will have, or at least act as if they have, "more general principles or rules of ethical conduct that are adjusted and interpreted to fit various kinds of situations," full ethical agents "can make ethical judgments about a wide variety of situations" and in many cases can "provide some justification for them."

Providing some examples of each, Moor notes that a "robotic camel jockey" (a technology used in Qatar to replace young boys as jockeys, thus freeing those boys from slavery in the human trafficking business) is an instance of an ethical impact agent. An airplane's automatic pilot system and an automatic teller machine (ATM) are both examples of an implicit ethical agent, since they have built-in programming designed to prevent harm from happening to the aircraft in one case and (in the other case) to prevent ATM customers from being shortchanged in financial transactions. Explicit ethical agents, on the other hand, would be able to calculate the best ethical action to take in a specific situation and would be able to make decisions when presented with ethical dilemmas. In Moor's scheme, full ethical agents have the kind of ethical features that we usually attribute to ethical agents like us (i.e., what Moor describes as "normal human adults"), including consciousness and free will.

Moor does not claim that either explicit or full ethical agents exist or that they will be available anytime in the near term. However, his distinctions are very helpful, as we try to understand various levels of moral agency that potentially affect AMs. Even if AMs may never qualify as full moral agents, Wallach and Allen (2009) believe that they can have "functional morality," based on two key criteria or dimensions: (i) autonomy and (ii) sensitivity to ethical values. However, Wallach and Allen also note that we do not yet have systems with both high autonomy and high sensitivity. They point out that an autopilot is an example of a system that has significant autonomy (in a limited domain) but little sensitivity to ethical values. On the contrary, the authors note that while ethical decision support systems (such as those used in the medical field to assist doctors) provide decision makers with access to morally relevant information and thus suggest high sensitivity to moral values, they have virtually no autonomy.

Wallach and Allen also argue that it is not necessary that AAs be moral agents in the sense that humans are. They believe that all we need to do is to design machines to act "as if" they are moral agents and thus "function" as such. We return to this point, as well as to the concept of functional morality, in Section 12.5.1. First, however, we ask if it makes sense to ascribe any level of morality, functional or otherwise, to AMs if those systems are not capable of being genuinely autonomous. While Wallach and Allen note that autonomy is one of the two key criteria in their framework of functional morality, they do not elaborate on the sense(s) in

which an AA can be said to be autonomous. We next examine the concept of autonomy to see whether an AM can indeed be autonomous.

Autonomy and "Functional Autonomy" in the Context of AMs

We briefly mentioned the concept of autonomy in Section 12.2.2 in our analysis of ethical concerns affecting AmI. There, we asked whether humans would, in effect, surrender some of their individual autonomy if they delegate (control of) certain kinds of tasks to computer systems. Some critics suggest that they might, especially if those computer systems are "autonomous." For example, Son (2015) notes that autonomous technologies can undermine "human autonomy" in ways that are both "subtle and indirect." Allen, Wallach, and Smit (2006), on the contrary, suggest that we need not worry about perceived threats to human autonomy because AMs will not necessarily "undermine our basic humanity." To evaluate these claims, however, we need a clear definition of *autonomy*.

Many philosophers associate autonomy with concepts such as liberty, dignity, and individuality.[18] Others, however, link autonomy to "independence." For example, O'Neill (2002) defines autonomy as a "capacity or trait that individuals manifest by acting independently." While it is difficult to ascribe characteristics such as liberty and dignity to AMs, we have seen that these machines do appear to be capable of "acting independently." So, if we can show that AMs can indeed act independently, it would seem plausible to describe AMs as entities that are also autonomous in some sense.

We should note that some influential definitions of autonomous systems and autonomous AAs link an artificial entity's ability to "adapt" to its environment with an ability to act "independently." For example, the Royal Academy's 2009 report seems to suggest that because autonomous systems are "adaptive," they also exhibit some degree of "independence." And Floridi (2008) makes a similar point, noting that an "adaptive" AA—one that can change its (internal) state dynamically, that is, without any external stimuli—has a certain degree of "*independence* from its environment."[19] Perhaps, then, AMs can satisfy O'Neill's requirement for autonomy by virtue of their capacity to act independently.

Insofar as AMs appear to be capable of acting independently, or behave "as if" they are acting independently, it would seem that we could attribute at least some degree of autonomy to them. Whether AMs will ever be capable of having full autonomy, in the sense that humans can, is debatable, and that question will not be examined here since it is beyond the scope of this chapter. However, an AM that can act independently in the sense described earlier can have "functional autonomy" and thus can qualify as a "functionally autonomous AM." We will next see that AMs must have some level of autonomy, even if only in a functional sense, if they are capable of being trusted by—that is, being in a trust relationship with—humans.

Trust and Authenticity in the Context of AMs

What does a relationship of trust involving humans and AMs entail? Lim, Stocker, and Larkin (2008) describe the possibility of a mutual or reciprocal trust relationship involving both (a) "Man to Machine" and (b) "Machine to Man." However, we limit our discussion to (a), and we ask two basic questions: (i) What would it mean for a human to *trust* an AM? (ii) Why is that question important? The significance of (ii) is highlighted in the Royal Academy of Engineering's report (2009), which asks whether we can trust AMs to always act in our best interests, especially AMs designed in such a way that they cannot be shut down by human operators. To answer (i), however, we first need to clarify what is meant by the concept of trust in general—that is, the kind of trust that applies in relationships between humans.

McLeod (2015) points out that trust, in human relationships, is both "important but dangerous." It is important because it enables us "to form relationships with others and to depend on them." But it is also dangerous, McLeod notes, because it involves risk. Since trusting someone "requires that we can be vulnerable to others (i.e., vulnerable to betrayal)," the trustor

(in the trust relationship) must be willing to accept some level of risk. In the case of AMs, we may be required to extend the level of risk beyond what we typically find acceptable for trust in human relationships. Before addressing that concern, however, it would be useful to establish what, exactly, is required for a normal trust relationship between humans.

A typical dictionary, such as the *American Heritage College Dictionary* (4th ed. 2002), defines trust as "firm reliance on the integrity, ability, or character of a person or thing." Definitions of trust that focus mainly on *reliance*, however, do not always help us to understand the nature of ethical trust. For example, I *rely* on my automobile engine to start today, but I do not "trust" it to do so. Conversely, I trust my daughter implicitly, but I cannot always rely on her to organize her important papers.[20] Thus, trust and reliance are not equivalent notions; while reliance may be a necessary condition for trust, something more is needed for ethical trust.

Because I am unable to have a trust relationship with a conventional machine such as an automobile, does it follow that I also cannot have one with an AM? Or does an AM's ability to exhibit some level of autonomy—even if only functional autonomy—make a difference? Consider that I am able to trust a human because the person in whom I place my trust not only can disappoint me (or let me down) but can also betray me—for example, that person, as a fully autonomous (human) agent, can freely elect to breach the trust I have placed in her. So it would seem that an entity's having at least some sense of autonomy is required for it to be capable of breaching the trust that someone has placed in it. In this sense, my automobile cannot breach my trust or betray me, even though I may be very disappointed if it fails to start today. Although my automobile does not have autonomy, we have seen that an AM has (functional) autonomy and thus might seem capable of satisfying the conditions required for a trust relationship. But even if an AM has (some level of) autonomy and even if having autonomy is a necessary condition for being in a trust relationship, it does not follow that it is a sufficient condition.[21] So, we can further ask whether any additional requirements may also need to be satisfied.

Some philosophers argue that trust has an emotive (or "affective") aspect and that this may be especially important in understanding trust in the context of AMs. For example, Coeckelbergh (2010) argues that if we want to build moral AMs (capable of trust), we will have to build them "*with emotions*."[22] Elsewhere, Coeckelbergh (2012) argues that for a trust relationship to be established between humans and machines, "appearance" (including the appearance of having emotions) is also very important. And because AMs may need to *appear* as if they have human-like properties, such as emotions, in order to be trusted by humans, we may be inclined to develop future AMs along these lines. Coeckelbergh and others seem to suggest that we should.[23]

Turkle (2011) raises some concerns involving emotions or feelings in the context of human–machine trust relationships, and she worries about what can happen when machines appear "as if" they have feelings. She describes a phenomenon called the "Eliza effect," which was initially associated with a response that some users had to an interactive software program called "Eliza" (designed by Joseph Weizenbaum at MIT in the 1960s). Turkle notes that this program, which was an early foray into machine learning programs designed to use language conversationally (and possibly pass the Turing test), solicited trust on the part of users. Eliza did this, Turkle points out, even though it was designed in a way that tricks users. Although Eliza was only a ("disembodied") software program, Turkle suggests that it could nevertheless be viewed as a "relational entity," or what she calls a "relational artifact," because of the way people responded to, and confided in, it. In this sense, Eliza seemed to have a strong emotional impact on some of the students who interacted with it. Turkle also notes that while Eliza "elicited trust" on the part of these students, it understood nothing about them.

Turkle worries that when a machine (as a relational artifact) appears to be interested in people, it can "push our Darwinian buttons . . . which causes people to respond *as if* they were in a relationship."[24] This is especially apparent in the case of physical AMs that are capable of facial expressions, such as Kismet (developed in MIT's AI Lab). Turkle suggests that because

AMs can be designed in ways that make people feel as if a machine cares about them (as in the case of Paro, a companion robot designed to comfort the elderly), people can develop feelings of trust in, and attachment to, that machine. For example, she notes that Cynthia Breazeal, one of Kismet's designers who had also developed a "maternal connection" with this AM while she was a student at MIT, had a difficult time separating from Kismet when she left that institution. In Turkle's view, this factor raises questions of both trust *and* authenticity, and Turkle worries that, unlike in the past, humans must now be able to distinguish between authentic and simulated relationships. While this connection between trust and authenticity/attachment opens up a new and provocative line of inquiry, and while it will be interesting to see how this connection eventually plays out in the context of trust and AMs, a further discussion of this topic would take us beyond the scope of this chapter.

In concluding this section, we note that many questions about trust vis-à-vis AMs have been left either unanswered or unexamined. Readers who are interested in learning more about this topic can consult the expanding literature on trust and e-trust in connection with artificial agents/entities.[25] Next, we ask how critical it is for humans to have a trust relationship with AMs as we pursue the goal of developing "moral machines." In other words, if we cannot trust AMs, should we build machines capable of making decisions that have significant moral impacts? And if not, do we need to reassess one of the core objectives of machine ethics?

▶ 12.5 MACHINE ETHICS AND *MORAL MACHINES*

The ethical issues examined in earlier chapters of this book arose mainly because of what we, as humans, do with computers and cybertechnology. In Section 12.4.2, however, we considered some AM-specific ethical concerns that arise because of what AMs are now capable of doing on their own. Increasingly, ethical concerns generated by the autonomous technologies/systems that comprise AMs are examined as issues in a relatively new subfield of cyberethics called *machine ethics* (Allen, Wendell, and Smit 2006; M. Anderson and S. Anderson 2011; and Moor 2006). Some, however, use the expression "robo-ethics" (Verrugio 2006; Decker and Gutmann 2012) or "robot ethics" (Capurro and Nagenborg 2009; Lin, Abney, and Bekey 2012) to describe the field that addresses these issues. Wallach and Allen (2009) note that other authors have also used expressions such as "agent ethics" and "bot ethics." However, we use "machine ethics" to include the wide range of ethical issues that arise in the context of AMs. And, as noted in Section 12.4, we use "autonomous machines" to refer to the cluster of autonomous technologies/systems that generate those ethical issues.

Analyzing the moral impacts of what AMs are capable of doing by themselves is one principal focus of machine ethics; we examined some of those impacts in Section 12.4.2. In this section, however, we briefly consider two very different kinds of questions affecting machine ethics: (i) What is the proper scope of this field, and what are its primary objectives? (ii) Is it possible to design "moral machines" (and if so, should we develop them)? We postpone our discussion of (ii) until Section 12.5.2 and begin with an analysis of (i).[26]

12.5.1 What is *Machine Ethics*?

Michael Anderson and Susan Leigh Anderson (2011) describe machine ethics as an interdisciplinary field of research that is primarily concerned with developing ethics for machines, as opposed to developing ethics for humans who "use machines." In their view, machine ethics is concerned with

> giving machines ethical principles, or a procedure for discovering ways to resolve ethical dilemmas they may encounter, enabling them to function in an ethically responsible manner through their own decision making.[27]

Susan Anderson (2011) points out that a central question in machine ethics is whether ethics is, or can be made, computable. She believes that it is and also suggests that it may be "prudent to begin to make ethics computable by first creating a program that acts as an ethical advisor to humans before attempting to build a full-fledged moral machine." We return to Anderson's suggestion at a later point in this section, in our discussion of how a prototype of a moral machine might initially function as an "ethical advisor" in a "dialogue" with humans.

Anderson draws some useful distinctions with regard to various levels at which machines could be designed to behave ethically. For our purposes, these can be organized into three levels, where a designer could:

a. Build "limitations" into a machine that would prevent it from causing moral harm

b. Embed an AM with instructions that would require it to behave in a particular way—that is, "according to an ideal ethical principle or principles that are *followed by the human designer*"

c. Embed an AM with "(an) ideal ethical principle(s) . . . and a learning procedure from which it can abstract (an) ideal ethical principle(s) in guiding its own actions"[28]

Whereas (a) represents the simplest design for ensuring that a machine behaves ethically, such a machine would seem capable of being only an "ethical impact agent" in James Moor's framework (described in Section 12.5.2). But a machine conforming to (b), on the other hand, would seem to qualify as an example of Moor's "implicit ethical agent." Anderson believes that machines built along the lines of (c) could conform to Moor's notion of "explicit ethical agent." She also believes that accomplishing (c) is the "ultimate goal" of machine ethics. In this case, an AM would be able not only to behave ethically but also be able to "justify its behavior" by expressing in "understandable language" the "intuitively acceptable ethical principle(s) that it has used to calculate its behavior."[29]

Wallach and Allen (2009) believe that one way in which the field of machine ethics has expanded upon traditional computer ethics is by asking *how* computers can be made into "explicit moral reasoners." In answering this question, Wallach and Allen first draw an important distinction between "reasoning about ethics" and "ethical decision making." For example, they acknowledge that even if one could build artificial systems capable of reasoning about ethics, it does not necessarily follow that these systems would be genuine "ethical decision makers." However, their main interest in how AMs can be made into moral reasoners is more practical than theoretical in nature, and they believe that the challenge of figuring out how to provide software/hardware agents with moral decision-making capabilities is urgent; in fact, they argue that the time to begin work on designing "moral machines" is now!

12.5.2 Designing *Moral Machines*

Can/should we build the kinds of moral machines that Wallach, Allen, and others urge us to develop? First, we can ask what is meant by the expression "*moral* machine." For example, are there "immoral machines"? Or are all machines simply amoral or non-moral, as many people tend to assume? The kind of moral machines that Wallach and Allen have in mind are AMs that are capable of both (i) making moral decisions and (ii) acting in ways that "humans generally consider to be ethically acceptable behavior." We should note that the idea of designing machines that could behave morally, that is, with a set of moral rules embedded in them, is not entirely new. In the 1940s, science fiction writer Isaac Asimov anticipated the need for ethical rules that would guide the robots of the future when he formulated his (now-classic) Three Laws of Robotics:

1. A robot may not injure a human being, or through inaction, allow a human being to come to harm.

2. A robot must obey orders given it by human beings except where such orders would 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.[30]

Numerous critics have questioned whether the three laws articulated by Asimov are adequate to meet the kinds of ethical challenges that current AMs pose. But relatively few of these critics have proposed clear and practical guidelines for how to embed machines with ethical instructions that would be generally acceptable to most humans. S. Anderson and M. Anderson (2011) and Wallach and Allen have each put forth some very thoughtful proposals for how this can be done. First, we consider Wallach and Allen's framework.

In describing how we can begin to build moral machines, Wallach and Allen point out that they are not interested in questions about developing a machine that is merely "instrumentally good." For example, a machine may be considered instrumentally good if it performs its tasks well. (Recall James Moor's distinction about computers as normative (non-moral) agents vs. moral agents, which we examined in Section 12.4.2.) Wallach and Allen are concerned with building moral machines, or what they also refer to as artificial moral agents (AMAs), that behave in ways that humans generally consider to be *morally good*. They point out, for example, that while Deep Blue is a good chess-playing system because it does well at chess (i.e., defeating the best human chess players), it cannot be viewed as a "good AMA" because it is not required to make the kinds of decisions that have moral import.

Wallach and Allen argue that a *good AMA* "can detect the possibility of human harm or neglect of duty, and can take steps to avoid or minimize the undesirable outcomes." But how, exactly, would such an AMA be designed? For example, which kinds of ethical reasoning procedures should we build into these systems—that is, should they be embedded with principles that favor utilitarian-like reasoning or deontology-like reasoning, or perhaps some combination of the two? Also, could the principles of virtue ethics be built into the software code embedded in these machines, if that were deemed to be essential or even desirable?

Embedding Ethical Theory/Reasoning Procedures into AMs

To appreciate the challenges involved in selecting the appropriate kind of ethical theory/reasoning to embed in AMs, Wallach and Allen consider how a computerized "driverless trolley" might react in the now classic scenario involving a "runaway trolley" (described in Chapter 2), where the "driver" (i.e., the AM) has to make a split-second decision (or calculation). Should the AM throw a switch that will cause the trolley to change tracks and (intentionally) run over one person who is standing on that track? Or should the AM do nothing, in which case the trolley will run over five people directly in its path? An AM designed to execute instructions compatible with utilitarian- or consequentialist-based reasoning would likely make a very different (moral) decision, or calculation, than one designed to execute code based on deontological reasoning.

Susan Anderson (2011) notes that the ethical theory of act utilitarianism (which, she believes, shows that ethics is indeed computable) is too "simplistic." She argues that this ethical theory, as well as theories based on absolute duties (e.g., in Kant's categorical imperative, described in Chapter 2), are not, in themselves at least, adequate to build into machines. Instead, she believes that an ethical theory similar to Ross' version of deontology (also described in Chapter 2), which provides the basis for what she calls a "prima facie duty approach," is more desirable. A virtue of Ross' theory, you may recall, is that it shows why it is often necessary to deliberate and weigh between duties when two or more of them conflict. But Anderson notes that a significant problem with Ross' theory is that it does not provide a clear mechanism or procedure for determining which duty overrides another in many situations where conflicts arise. So, she supplements the prima facie duty approach with a "decision

principle" to resolve the conflicts that will inevitably arise. Anderson further argues that the kinds of "decision principles" needed to accomplish this "can be discovered by a machine"—that is, a machine could use an "inductive logic program" to arrive at such a principle. For example, Anderson believes that the machine could "learn from generalizing correct answers in particular cases."

Earlier in this section, we briefly mentioned Anderson's suggestion that it would be prudent for us first to design an artificial system to function as an "ethical advisor" to humans before attempting to build a full-fledged moral machine. Along similar lines, Susan Anderson and Michael Anderson (2011) have recommended building artificial systems with which humans can have an "ethical dialogue" before we embed machines themselves with ethical reasoning algorithms that they could use in a fully independent manner. The Andersons have developed such a system—that is, an "automated dialogue"—involving an ethicist and an artificial system that functions "more or less independently in a *particular domain*." They believe that this is an important first step in building moral machines because it enables the artificial system to learn both (i) the "ethically relevant features of the dilemmas it will encounter" (within that domain) and (ii) the appropriate prima facie duties and decision principles it will need to resolve the dilemmas.[31]

Functional Morality and a "Moral Turing Test"

Earlier in this section, we asked whether AMs are capable, in principle, of being genuine moral agents. Recall our brief discussion of Wallach and Allen's notion of functional morality, which the authors contrast with mere "operational morality" as well as with full moral agency. Wallach and Allen argue that even if machines fail to achieve full-blown moral agency, they may exhibit varying degrees of functional morality. So, they leave open the question of whether AMs could ever be full moral agents. Perhaps a more basic question to consider, however, is whether we could ever conclusively determine that we had developed an AM that was a full moral agent. Allen, Varner, and Zinser (2000) consider how a "Turing-like" test, which they call a "moral Turing test" (MTT), could be applied in response to this question.

Unlike the original Turing test (described in Chapter 11), the MTT shifts the focus in the human–machine interaction away from an emphasis on mere "conversational ability" to criteria involving questions about "action." In this case, an AM would be asked questions about how it would act in such and such a situation, as opposed to being evaluated in terms of how successfully it was able to converse with humans about topics involving moral principles and rules. However, Allen et al. reported that they still encountered several problems with this test as a procedure for conclusively establishing whether AMs could, in principle, qualify as full moral agents.

We have seen that Wallach and Allen seem far less concerned with questions about whether AMs can be full moral agents than with questions about how we can design AMs to act in ways that conform to our received notions of morally acceptable behavior. And Susan Anderson (2011) echoes this point when she notes that her primary concern also is with whether machines "can perform morally correct actions and can justify them if asked." We should note that Wallach and Allen also believe that questions about whether AMs can be full moral agents can actually distract from (what they consider to be) the more "important question about how to design systems to act appropriately in morally charged situations."[32]

Acknowledging that many important questions in machine ethics remain unresolved, we conclude this section by briefly identifying some reasons why continued work in machine ethics is important. Moor (2006) offers three such reasons: (i) Ethics (itself) is important, (ii) future machines will likely have increased autonomy, and (iii) designing machines to behave ethically will help us better understand ethics. Moor's third point ties in nicely with Wallach and Allen's claim that developments in machine ethics could help us to better understand our own nature as moral reasoners. In fact, they believe that research and development in machine ethics can

provide feedback for "humans' understanding of themselves as moral agents" and for our understanding of "the nature of ethical thinking itself."[33]

▶ 12.6 A "DYNAMIC" ETHICAL FRAMEWORK FOR GUIDING RESEARCH IN NEW AND EMERGING TECHNOLOGIES

We have considered a fairly wide range of ethical concerns affecting the new and emerging technologies examined in this chapter. Some of these ethical concerns directly impact the software engineers and computer professionals who design the technologies. But virtually everyone will be affected by these technologies in the near future; so all of us would benefit from clear policies and ethical guidelines that address research/development in new and emerging technologies. Moor (2008) argues that because these technologies promise "dramatic change," it is no longer satisfactory to do "ethics as usual." He goes on to claim that we need to be better informed in our "ethical thinking" and more proactive in our "ethical action."

What kind of ethical framework will we need to address the specific challenges posed by new and emerging technologies? One requirement would seem that this framework be "proactive" in its approach to ethics, as Moor suggests. Perhaps, then, we could look to the now classic *ELSI* (Ethical/Legal/Social Issues) model for some guidance on how to construct a proactive ethical framework for other emerging technologies as well. This model, which was initially developed for the Human Genome Project (HGP), was designed to anticipate the kinds of ethical, legal, and social implications that would likely arise in HGP research. Before work on HGP was allowed to proceed, the National Human Genome Research Institute (NHGRI) required that ethical, legal, and social issues first had to be identified and addressed. Many of the salient features of the original ELSI model for HGP—requirements that addressed concerns affecting privacy, confidentiality, fairness, etc.—were "built into" the scientific research methodology used for HGP.[34]

12.6.1 Is an ELSI-Like Model Adequate for New/Emerging Technologies?

Should the original ELSI model, or one similar to it, be used to guide the development of other new/emerging technologies as well? ELSI's proponents believe that it is an ideal model because it is, as we noted, "proactive." They point out that prior to the ELSI program, ethics was typically "reactive" in the sense that it "followed scientific developments" rather than informing scientific research. As Moor and others note, ethics has had to play "catch up" in most scientific research areas because ethical guidelines were developed in response to cases where serious harm had already resulted. For these reasons, Kurzweil (2005) believes that a proactive ethical framework is needed in nanotechnology research, and he has suggested that an ELSI-like model be developed to guide researchers working in that technological field. The Royal Academy of Engineering's influential report on autonomous systems (2009) has also suggested an ELSI-like framework be used to assess ethical, legal, and social issues that affect or will soon affect autonomous technologies now under development.

Although many see ELSI as a vast improvement over traditional frameworks, the standard ELSI model employs a scheme that Moor and Weckert (2004) describe as an "ethics-first" framework. They believe that ethical frameworks of this kind have problems because they depend, in large part, on a "factual determination" of the specific harms and benefits in implementing the technology before an ethical assessment can be done. But the authors note that in the case of nanotechnology developments, for example, it is very difficult to know what the future will be in 5 or 10 years, let alone 20 or more years. So if we adopt an ("ethics-first") ELSI-like model, it might seem appropriate to put a moratorium on research in an area of technology until we get all of the facts. However, Moor and Weckert point out that while a

moratorium on future research would halt technology developments in a field, such as nanotechnology for example, it will not advance ethics in that technological area.

12.6.2 A "Dynamic Ethics" Model

Moor and Weckert also argue that turning back to what they call the "ethics-last model" is not desirable either. The authors note that once a technology is in place, much unnecessary harm may already have occurred. So, in Moor and Weckert's scheme, neither an ethics-first nor an ethics-last model is satisfactory for emerging technologies. In their view, ethics is something that needs to be done *continually* as a technology develops and as its "potential social consequences become better understood." The authors also point out that ethics is "dynamic" in the sense that the factual/descriptive component on which the normative analysis relies has to be continually updated.

As we debate whether to go forward with research and development in a particular new or emerging technology, we can see how neither an ethics-first nor an ethics-last model is adequate. We can also agree with Moor and Weckert that it is necessary to establish a set of ethical criteria that can be continually updated as new factual information about that technology becomes available. This point needs to be specified in any viable ethical framework, as well as in any effective set of policy guidelines, that we implement.

Recall the comprehensive cyberethics framework that we articulated at the end of Chapter 1, which included three steps: (i) *identify* a controversial issue (*or* practice *or* technological feature) involving cybertechnology, (ii) *analyze* the ethical issue(s) involved by clarifying relevant concepts, and (iii) *deliberate* on the ethical issue(s) in terms of one or more standard ethical theories (e.g., utilitarianism, deontology, etc.). Building on Moor and Weckert's insights regarding ethical challenges posed by new and emerging technologies, we add a fourth component or step to that framework:

(iv) Update the ethical analysis by continuing to:

a. Differentiate between the factual/descriptive and normative components of the new or emerging technology under consideration

b. Revise the policies affecting that technology as necessary, especially as the factual data or components change or as information about the potential social impacts becomes clearer

As information about plans for the design and development of a new technology becomes available, we can loop back to (i) and proceed carefully through each step in the expanded ethical framework. This four-step framework can also be applied as new information about existing technologies and their features becomes available.

▶ 12.7 CHAPTER SUMMARY

In this chapter, we examined a cluster of ethical and social challenges affecting emerging and converging technologies. In particular, we described and evaluated controversies involving two broad areas of technological convergence: AmI and nanocomputing. We saw that AmI environments, made possible by pervasive computing and ubiquitous communication, raised concerns for freedom and autonomy as well as for privacy and surveillance. We also saw some ways in which ongoing developments in nanotechnology will likely raise concerns regarding privacy, longevity, and "runaway nanobots." We then examined some ethical challenges posed by "autonomous machines," and we considered whether it might be possible to design "moral machines." Finally, we argued that a "dynamic" ethical framework, introduced by James Moor