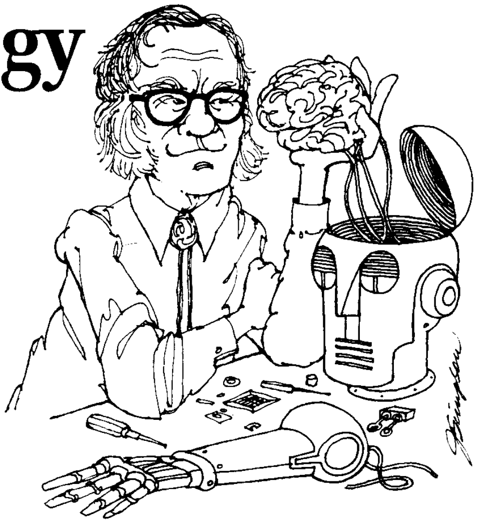
COMPUTING MILIEUX

Asimov's Laws of Robotics: Implications for Information Technology

Part 1



Roger Clarke

Australian National University

[redacted]ith the death of Isaac Asimov on April 6, 1992, the world lost a prodigious imagination. Unlike fiction writers before him, who regarded robotics as something to be feared, Asimov saw a promising technological innovation to be exploited and managed. Indeed, Asimov's stories are experiments with the enormous potential of information technology.

This article examines Asimov's stories not as literature but as a *gedankenexperiment* — an exercise in thinking-through the ramifications of a design. Asimov's intent was to devise a set of rules that would provide reliable control over semiautonomous machines. My goal is to determine whether such an achievement is likely or even possible in the real world. In the process, I focus on practical, legal, and ethical matters that may have short- or medium-term implications for practicing information technologists.

Part 1, in this issue, reviews the origins of the robot notion and explains the laws for controlling robotic behavior, as espoused by Asimov in 1940 and presented and refined in his writings over the following 45 years. Next month, Part 2 examines the implications of Asimov's fiction not only for real roboticists but also for information technologists in general.

Because many contemporary applications of information technology exhibit robotic characteristics, the difficulties Isaac Asimov identified in his stories are directly relevant to information technology professionals.

Origins of robotics

Robotics, a branch of engineering, is also a popular source of inspiration in science fiction literature; indeed, the term originated in that field. Many authors have written about robot behavior and their interaction with humans, but in this company Isaac Asimov stands supreme. He entered the field early, and from 1940 to 1990 he dominated it. Most subsequent science fiction literature expressly or implicitly recognizes his Laws of Robotics.

0018-9162/93/1200-0053 $03.00 © 1993 IEEE

Asimov described how, at the age of 20, he came to write robot stories:

> In the 1920s science fiction was becoming a popular art form for the first time... and one of the stock plots... was that of the invention of a robot.... Under the influence of the well-known deeds and ultimate fate of Frankenstein and Rossum, there seem-ed only one change to be rung on this plot — robots were created and destroyed their creator....I quickly grew tired of this dull hundred-times-told tale....Knowledge has its dangers, yes, but is the response to be a retreat from knowledge?... I began, in 1940, to write robot stories of my own — but robot stories of a new variety.... My robots were machines designed by engineers, not pseudomen created by blasphemers.[1,2]

Asimov was not the first to conceive of well-engineered, nonthreatening robots, but he pursued the theme with such enormous imagination and persistence that most of the ideas that have emerged in this branch of science fiction are identifiable in his stories.

To cope with the potential for robots to harm people, Asimov, in 1940, in conjunction with science fiction author and editor John W. Campbell, formulated the Laws of Robotics.[3,4] He subjected all of his fictional robots to these laws by having them incorporated within the architecture of their (fictional) "platinum-iridium positronic brains." The laws (see sidebar on next page) first appeared publicly in his fourth robot short story, "Runaround."[5]

The laws quickly attracted — and have since retained — the attention of readers and other science fiction writers. Only two years later another established writer, Lester Del Rey, referred to "the mandatory form that would force built-in unquestioning obedience from the robot."[6] As Asimov later wrote (with his characteristic clarity and lack of modesty), "Many writers of robot stories, without actually quoting the three laws, take them for granted, and expect the readers to do the same."[1]

Asimov's fiction even influenced the origins of robotic engineering. "Engelberger, who built the first indus trial robot, called Unimate, in 1958, attributes his long-standing fascination with robots to his reading of [Asimov's] 'I, Robot' when he was a teenager,"[4] and Engelberger later invited Asimov to write the foreword to his robotics manual.

The laws are intuitively appealing: They are simple and straightforward, and they embrace "the essential guiding principles of a good many of the world's ethical systems."[7] They also appear to ensure the continued dominion of humans over robots, and to preclude the use of robots for evil purposes. In practice, however — meaning in Asimov's numerous and highly imaginative stories — a variety of difficulties arise.

My purpose here is to determine whether or not Asimov's fiction vindicates the laws he expounded. Does he successfully demonstrate that robotic technology can be applied in a responsible manner to potentially powerful, semiautonomous, and, in some sense, intelligent machines? To reach a conclusion, we must examine many issues emerging from Asimov's fiction.

History. The robot notion derives from two strands of thought, humanoids and automata. The notion of a humanoid (or human-like nonhuman) dates back to Pandora in *The Iliad*, 2,500 years ago—and even further. Egyptian, Babylonian, and ultimately Sumerian legends fully 5,000 years old reflect the widespread image of the creation, with god-men breathing life into clay models. One variation on the theme is the idea of the golem, associated with the Prague ghetto of the sixteenth century. This clay model, when breathed into life, became a useful but destructive ally.

The golem was an important precursor to Mary Shelley's *Frankenstein: The Modern Prometheus* (1818). This story combined the notion of the humanoid with the dangers of science (as suggested by the myth of Prometheus, who stole fire from the gods to give it to mortals). In addition to establishing a literary tradition and the genre of horror stories, *Frankenstein* also imbued humanoids with an aura of ill fate.

Automata, the second strand of thought, are literally "self-moving things" and have long interested mankind. Early models depended on levers and wheels, or on hydraulics. Clockwork technology enabled significant advances after the thirteenth century, and later steam and electro-

Isaac Asimov, 1920-1992

Born near Smolensk in Russia, Isaac Asimov came to the United States with his parents three years later. He grew up in Brooklyn, becoming a US citizen at the age of eight. He earned bachelor's, master's, and doctoral degrees in chemistry from Columbia University and qualified as an instructor in biochemistry at Boston University School of Medicine, where he taught for many years and performed research in nucleic acid.

As a child, Asimov had begun reading the science fiction stories on the racks in his family's candy store, and those early years of vicarious visits to strange worlds had filled him with an undying desire to write his own adventure tales. He sold his first short story in 1938, and after wartime service as a chemist and a short hitch in the Army, he focused increasingly on his writing.

Asimov was among the most prolific of authors, publishing hundreds of books on various subjects and dozens of short stories. His Laws of Robotics underlie four of his full-length novels as well as many of his short stories. The World Science Fiction Convention bestowed Hugo Awards on Asimov in nearly every category of science fiction, and his short story "Nightfall" is often referred to as the best science fiction story ever written. The scientific authority behind his writing gave his stories a feeling of authenticity, and his work undoubtedly did much to popularize science for the reading public.

mechanics were also applied. The primary purpose of automata was entertainment rather than employment as useful artifacts. Although many patterns were used, the human form always excited the greatest fascination. During the twentieth century, several new technologies moved automata into the utilitarian realm. Geduld and Gottesman[8] and Frude[2] review the chronology of clay model, water clock, golem, homunculus, android, and cyborg that culminated in the contemporary concept of the robot.

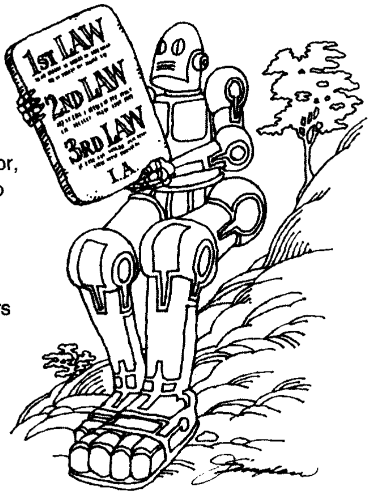
The term robot derives from the Czech word *robota*, meaning forced work or compulsory service, or *robotnik*, meaning serf. It was first used by the Czech playwright Karel Çapek in 1918 in a short story and again in his 1921 play *R.U.R.*, which stood for Rossum's Universal Robots. Rossum, a fictional Englishman, used biological methods to invent and mass-produce "men" to serve humans. Eventually they rebelled, became the dominant race, and wiped out humanity. The play was soon well known in English-speaking countries.

Asimov's Laws of Robotics (1940)

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

Second Law:
A robot must obey the orders given it by human beings, except where such orders would conflict with the First Law.

Third Law:
A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.



Definition. Undeterred by its somewhat chilling origins (or perhaps ignorant of them), technologists of the 1950s appropriated the term robot to refer to machines controlled by programs. A robot is "a reprogrammable multifunctional device designed to manipulate and/or transport material through variable programmed motions for the performance of a variety of tasks."[9] The term robotics — which Asimov claims he coined in 1942[10] — refers to "a science or art involving both artificial intelligence (to reason) and mechanical engineering (to perform physical acts suggested by reason)."[11]

As currently defined, robots exhibit three key elements:

- programmability, implying computational or symbol-manipulative capabilities that a designer can combine as desired (a robot is a computer);
- mechanical capability, enabling it to act on its environment rather than merely function as a data processing or computational device (a robot is a machine); and
- flexibility, in that it can operate using a range of programs and manipulate and transport materials in a variety of ways.

We can conceive of a robot, therefore, as either a computer-enhanced machine or as a computer with sophisticated input/output devices. Its computing capabilities enable it to use its motor devices to respond to external stimuli, which it detects with its sensory devices. The responses are more complex than would be possible using mechanical, electromechanical, and/or electronic components alone.

With the merging of computers, telecommunications networks, robotics, and distributed systems software, and the multiorganizational application of the hybrid technology, the distinction between computers and robots may become increasingly arbitrary. In some cases it would be more convenient to conceive of a principal intelligence with dispersed sensors and effectors, each with subsidiary intelligence (a robotics-enhanced computer system). In others would be more realistic to think in terms of multiple devices, each with appropriate sensory, processing, and motor capabilities, all subjected to some form of coordination (an integrated multirobot system). The key difference robotics brings is the complexity and persistence that artifact behavior achieves, independent of human involvement.

Many industrial robots resemble humans in some ways. In science fiction the tendency has been even more pronounced, and readers encounter humanoid robots, humaniform robots, and androids. In fiction, as in life, it appears that a robot needs to exhibit only a few human-like characteristics to be treated as if it were human. For example, the relationships between humans and robots in many of Asimov's stories seem almost intimate, and audiences worldwide reacted warmly to the "personality" of the computer HAL in *2001: A Space Odyssey*, and to the gibbering rubbish-bin R2-D2 in the Star Wars series.

The tendency to conceive of robots in humankind's own image may gradually yield to utilitarian considerations, since artifacts can be readily designed to transcend humans' puny sensory and motor capabilities. Frequently the disadvantages and risks involved in incorporating sensory, processing, and motor apparatus within a single housing clearly outweigh the advantages.

Many robots will therefore be anything but humanoid in form. They may increasingly comprise powerful processing capabilities and associated memories in a safe and stable location, communicating with one or more sensory and motor devices (supported by limited computing capabilities and memory) at or near the location(s) where the robot performs its functions. Science fiction literature describes such architectures.[12,13]

Impact. Robotics offers benefits such as high reliability, accuracy, and speed of operation. Low long-term costs of computerized machines may result in significantly higher productivity, particularly in work involving variability within a general pattern. Humans can be relieved of mundane work and exposure to dangerous workplaces. Their capabilities can be extended into hostile environments involving high pressure (deep water), low pressure (space), high temperatures (furnaces), low temperatures (ice caps and cryogenics), and high-radiation areas (near nuclear materials or occurring naturally in space).

On the other hand, deleterious consequences are possible. Robots might directly or indirectly harm humans or their property; or the damage may be economic or incorporeal (for example, to a person's reputation). The harm could be accidental or result from human instructions. Indirect harm may occur to workers, since the application of robots generally results in job redefinition and sometimes in outright job displacement. Moreover, the replacement of humans by machines may undermine the self-respect of those affected, and perhaps of people generally.

During the 1980s, the scope of information technology applications and their impact on people increased dramatically. Control systems for chemical processes and air conditioning are examples of systems that already act directly and powerfully on their environments. And consider computer-integrated manufacturing, just-in-time logistics, and automated warehousing systems. Even data processing systems have become integrated into organizations' operations and constrain the ability of operations-level staff to query a machine's decisions and conclusions. In short, many modern computer systems are arguably robotic in nature already; their impact must be managed — now.

The 1940 Laws of Robotics

Asimov's original laws (see previous sidebar) provide that robots are to be slaves to humans (the second law). However, this role is overridden by the higher order first law, which precludes robots from injuring a human, either by their own autonomous action or by following a human's instructions. This precludes their continuing with a programmed activity when doing so would result in human injury. It also prevents their being used as a tool or accomplice in battery, murder, self-mutilation, or suicide.

The third and lowest level law creates a robotic survival instinct. This ensures that, in the absence of conflict with a higher order law, a robot will

- seek to avoid its own destruction through natural causes or accident,
- defend itself against attack by another robot or robots, and
- defend itself against attack by any human or humans.

Being neither omniscient nor omnipotent, it may of course fail in its endeavors. Moreover, the first law ensures that the robotic survival instinct fails if self-defense would necessarily involve injury to any human. For robots to successfully defend themselves against humans, they would have to be provided with sufficient speed and dexterity so as not to impose injurious force on a human.

Under the second law, a robot appears to be required to comply with a human order to (1) not resist being destroyed or dismantled, (2) cause itself to be destroyed, or (3) (within the limits of paradox) dismantle itself.[12] In various stories, Asimov notes that the order to self-destruct does not have to be obeyed if obedience would result in harm to a human. In addition, a robot would generally not be precluded from seeking clarification of the order. In his last full-length novel, Asimov appears to go further by envisaging that court procedures would be generally necessary before a robot could be destroyed: "I believe you should be dismantled without delay. The case is too dangerous to await the slow majesty of the law. . . . If there are legal repercussions hereafter, I shall deal with them."[14]

Such apparent inconsistencies attest to the laws' primary role as a literary device intended to support a series of stories about robot behavior. In this, they were very successful: "There was just enough ambiguity in the Three Laws to provide the conflicts and uncertainties required for new stories, and, to my great relief, it seemed always to be possible to think up a new angle out of the 61 words of the Three Laws."[1]

As Frude says, "The Laws have an interesting status. They . . . may easily be broken, just as the laws of a country may be transgressed. But Asimov's provision for building a representation of the Laws into the positronic-brain circuitry ensures that robots are physically prevented from contravening them."[2] Because the laws are intrinsic to the machine's design, it should "never even enter into a robot's mind" to break them.

Subjecting the laws to analysis may seem unfair to Asimov. However, they have attained such a currency not only among sci-fi fans but also among practicing roboticists and software developers that they influence, if only subconsciously, the course of robotics.

Asimov's experiments with the 1940 laws

Asimov's early stories are examined here not in chronological sequence or on the basis of literary devices, but by looking at clusters of related ideas.

The ambiguity and cultural dependence of terms. Any set of "machine values" provides enormous scope for linguistic ambiguity. A robot must be able to distinguish robots from humans. It must be able to recognize an order and distinguish it from a casual request. It must "understand" the concept of its own existence, a capability that arguably has eluded mankind, although it may be a simpler matter for robots. In one short story, for example, the vagueness of the word *firmly* in the order "Pull [the bar] towards you firmly" jeopardizes a vital hyperspace experiment. Because robot strength is much greater than that of humans, it pulls the bar more powerfully than the human had intended, bends it, and thereby ruins the control mechanism.[15]

Defining injury and harm is particularly problematic, as are distinctions between death, mortal danger, and injury or harm that is not life-threatening. Beyond this, there is psychological harm. Any robot given, or developing, an awareness of human feelings would have to evaluate injury and harm in psychological as well as physical terms: "The insurmountable First Law of Robotics states: 'A robot may not injure a human being . . .' , and *to repel a friendly gesture would do injury*"[16] (emphasis added). Asimov investigated this in an early short story and later in a novel: A mind-reading robot interprets the first law as requiring him to give people not the correct answers to their questions but the answers that he knows they want to hear.[14,16,17]

Another critical question is how a robot is to interpret the term human. A robot could be given any number of subtly different descriptions of a human being, based, for example, on skin color, height range, and/or voice characteristics such as accent. It is therefore possible for robot behavior to be manipulated: "The Laws, even the First Law, might not be an absolute then, but might be whatever those who design robots define them to be."[14] Faced with this difficulty, the robots in this story conclude that "... if different robots are subject to narrow definitions of one sort or another, there can only be measureless destruction. We define human beings as all members of the species, Homo sapiens."[14]

In an early short story, Asimov has a humanoid robot represent itself as a human and stand for public office. It must prevent the public from realizing that it is a robot, since public reaction would not only result in its losing the election but also in tighter constraints on other robots. A political opponent, seeking to expose the robot, discovers that it is impossible to prove it is a robot solely on the basis of its behavior, because the Laws of Robotics force any robot to perform in essentially the same manner as a good human being.[7] In a later novel, a roboticist says, "If a robot is human enough, he would be accepted as human. Do you demand proof that I am not a robot? The fact that I *seem* human is enough."[16] In another scene, a humaniform robot is sufficiently similar to a human to confuse a normal robot and slow down its reaction time.[14] Ultimately, two advanced robots recognize each other as "human," at least for the purposes of the laws.[14,18]

Does the prosthetization of humans lead inevitably to the humanization of robots?

Defining human beings becomes more difficult with the emergence of cyborgs, which may be seen as either machine-enhanced humans or biologically enhanced machines. When a human is augmented by prostheses (artificial limbs, heart pacemakers, renal dialysis machines, artificial lungs, and someday perhaps many other devices), does the notion of a human gradually blur with that of a robot? And does a robot that attains increasingly human characteristics (for example, a knowledge-based system provided with the "know-that" and "know-how" of a human expert and the ability to learn more about a domain) gradually become confused with a human? How would a robot interpret the first and second laws once the Turing test criteria can be routinely satisfied? The key outcome of the most important of Asimov's robot novellas[12] is the tenability of the argument that the prosthetization of humans leads inevitably to the humanization of robots.

The cultural dependence of meaning reflects human differences in such matters as religion, nationality, and social status. As robots become more capable, however, cultural differences between humans and robots might also be a factor. For example, in one story[19] a human suggests that some laws may be bad and their enforcement unjust, but the robot replies that an unjust law is a contradiction in terms. When the human refers to something higher than justice, for example, mercy and forgiveness, the robot merely responds, "I am not acquainted with those words."

The role of judgment in decision making. The assumption that there is a literal meaning for any given series of signals is currently considered naive. Typically, the meaning of a term is seen to depend not only on the context in which it was originally expressed but also on the context in which it is read (see, for example, Winograd and Flores[20]). If this is so, then robots must exercise judgment to interpret the meanings of words and hence of orders and of new data.

A robot must even determine whether and to what extent the laws apply to a particular situation. Often in the robot stories a robot action of any kind is impossible without some degree of risk to a human. To be at all useful to its human masters, a robot must therefore be able to judge how much the laws can be breached to maintain a tolerable level of risk. For example, in Asimov's very first robot short story, "Robbie [the robot] snatched up Gloria [his young human owner], slackening his speed not one iota, and, consequently, knocking every breath of air out of her."[21] Robbie judged that it was less harmful for Gloria to be momentarily breathless than to be mown down by a tractor.

Similarly, conflicting orders may have to be prioritized, for example, when two humans give inconsistent instructions. Whether the conflict is overt, unintentional, or even unwitting, it nonetheless requires a resolution. Even in the absence of conflicting orders, a robot may need to recognize foolish or illegal orders and decline to implement them, or at least question them. One story asks, "Must a robot follow the orders of a child; or of an idiot; or of a criminal; or of a perfectly decent intelligent man who happens to be inexpert and therefore ignorant of the undesirable consequences of his order?"[18]

Numerous problems surround the valuation of individual humans. First, do all humans have equal standing in a robot's evaluation? On the one hand they do: "A robot may not judge whether a human being deserves death. It is not for him to decide. He may not harm a human — variety skunk or variety angel."[7] On the other hand they might not, as when a robot tells a human, "In conflict between your safety and that of another, I must guard yours."[22] In another short story, robots agree that they "must obey a human being who is fit by mind, character, and knowledge to give me that order." Ultimately, this leads the robot to "disregard shape and form in judging between human beings" and to recognize his companion robot not merely as human but as a human "more fit than the others."[18] Many subtle problems can be constructed. For example, a person might try forcing a robot to comply with an instruction to harm a human (and thereby violate the first law) by threatening to kill himself unless the robot obeys.

How is a robot to judge the trade-off between a high probability of lesser harm to one person versus a low probability of more serious harm to another? Asimov's stories refer to this issue but are somewhat inconsistent with each other and with the strict wording of the first law.

More serious difficulties arise in relation to the valuation of multiple humans. The first law does not even contemplate the simple case of a single terrorist threatening many lives. In a variety of stories, however, Asimov interprets the law to recognize circumstances in which a robot may have to injure or even kill one or more humans to protect one or more others: "The Machine cannot harm a human being more than minimally, and that only to save *a greater number*"[23] (emphasis added). And again: "The First Law is not absolute. What if harming a human being saves the lives of two others, or three others, or even three billion others? The robot may have thought that saving the Federation took precedence over the saving of one life."[24]

The more subtle life-and-death cases might fall well outside a robot's appreciation.

These passages value humans exclusively on the basis of numbers. A later story includes this justification: "To expect robots to make judgments of fine points such as talent, intelligence, the general usefulness to society, has always seemed impractical. That would delay decision to the point where the robot is effectively immobilized. So we go by numbers."[18]

A robot's cognitive powers might be sufficient for distinguishing between attacker and attackee, but the first law alone does not provide a robot with the means to distinguish between a "good" person and a "bad" one. Hence, a robot may have to constrain a "good" attackee's self-defense to protect the "bad" attacker from harm. Similarly, disciplining children and prisoners may be difficult under the laws, which would limit robots' usefulness for supervision within nurseries and penal institutions.[22] Only after many generations of self-development does a humanoid robot learn to reason that "what seemed like cruelty [to a human] might, in the long run, be kindness."[12]

The more subtle life-and-death cases, such as assistance in the voluntary euthanasia of a fatally ill or injured person to gain immediate access to organs that would save several other lives, might fall well outside a robot's appreciation. Thus, the first law would require a robot to protect the threatened human, unless it was able to judge the steps taken to be the least harmful strategy. The practical solution to such difficult moral questions would be to keep robots out of the operating theater.[22]

The problem underlying all of these issues is that most probabilities used as input to normative decision models are not objective; rather, they are estimates of probability based on human (or robot) judgment. The extent to which judgment is central to robotic behavior is summed up in the cynical rephrasing of the first law by the major (human) character in the four novels: "A robot must not hurt a human being, unless he can think of a way to prove it is for the human being's ultimate good after all."[19]

The sheer complexity. To cope with the judgmental element in robot decision making, Asimov's later novels introduced a further complication: "On . . . [worlds other than Earth], . . . the Third Law is distinctly stronger in comparison to the Second Law. . . . An order for self-destruction would be questioned and there would have to be a truly legitimate reason for it to be carried through — a clear and present danger."[16] And again, "Harm through an active deed outweighs, in general, harm through passivity — all things being reasonably equal. . . . [A robot is] always to choose truth over nontruth, if the harm is roughly equal in both directions. In general, that is."[16]

The laws are not absolutes, and their

force varies with the individual machine's programming, the circumstances, the robot's previous instructions, and its experience. To cope with the inevitable logical complexities, a human would require not only a predisposition to rigorous reasoning, and a considerable education, but also a great deal of concentration and composure. (Alternatively, of course, the human may find it easier to defer to a robot suitably equipped for fuzzy-reasoning-based judgment.)

The strategies as well as the environmental variables involve complexity. "You must not think . . . that robotic response is a simple yes or no, up or down, in or out. . . . There is the matter of speed of response."[16] In some cases (for example, when a human must be physically restrained), the degree of strength to be applied must also be chosen.

The scope for dilemma and deadlock. A deadlock problem was the key feature of the short story in which Asimov first introduced the laws. He constructed the type of stand-off commonly referred to as the "Buridan's ass" problem. It involved a balance between a strong third-law self-protection tendency, causing the robot to try to avoid a source of danger, and a weak second-law order to approach that danger. "The conflict between the various rules is [meant to be] ironed out by the different positronic potentials in the brain," but in this case the robot "follows a circle around [the source of danger], staying on the locus of all points of . . . equilibrium."[5]

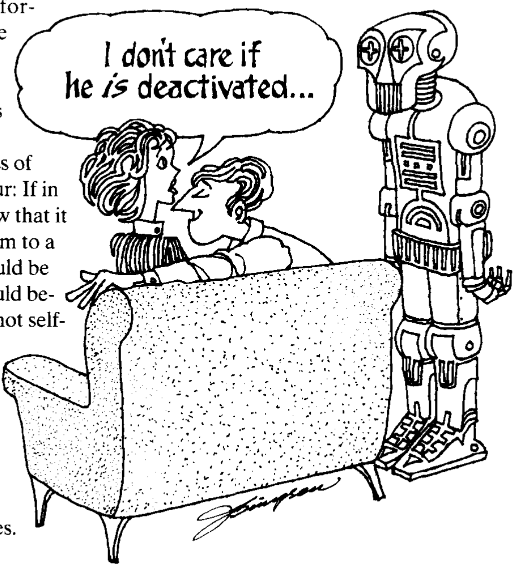
Deadlock is also possible within a single law. An example under the first law would be two humans threatened with equal danger and the robot unable to contrive a strategy to protect one without sacrificing the other. Under the second law, two humans might give contradictory orders of equivalent force. The later novels address this question with greater sophistication:

> What was troubling the robot was what roboticists called an equipotential of contradiction on the second level. Obedience was the Second Law and [the robot] was suffering from two roughly equal and contradictory orders. Robot-block was what the general population called it or, more frequently, roblock for short . . . [or] "mental freeze-out." . . . No matter how subtle and intricate a brain might be, there is always some way of setting up a contradiction. This is a fundamental truth of mathematics.[16]

Clearly, robots subject to such laws need to be programmed to recognize deadlock and either choose arbitrarily among the alternative strategies or arbitrarily modify an arbitrarily chosen strategy variable (say, move a short distance in any direction) and reevaluate the situation: "If A and not-A are precisely equal misery-producers according to his judgment, he chooses one or the other in a completely unpredictable way and then follows that unquestioningly. He does *not* go into mental freeze-out."[16]

The finite time that even robot decision making requires could cause another type of deadlock. Should a robot act immediately, by "instinct," to protect a human in danger? Or should it pause long enough to more carefully analyze available data — or collect more data — perhaps thereby discovering a better solution, or detecting that other humans are in even greater danger? Such situations can be approached using the techniques of information economics, but there is inherent scope for ineffectiveness and deadlock, colloquially referred to as "paralysis by analysis."

Asimov suggested one class of deadlock that would not occur: If in a given situation a robot knew that it was powerless to prevent harm to a human, then the first law would be inoperative; the third law would become relevant, and it would not self-immolate in a vain attempt to save the human.[25] It does seem, however, that the deadlock is not avoided by the laws themselves, but rather by the presumed sophistication of the robot's decision-analytical capabilities.



A special case of deadlock arises when a robot is ordered to wait. For example, "'[Robot], you will not move nor speak nor hear us until I say your name again.' There was no answer. The robot sat as though it were cast out of one piece of metal, and it would stay so until it heard its name again."[26] As written, the passage raises the intriguing question of whether passive hearing is possible without active listening. What if the robot's name is next used in the third person rather than the second?

In interpreting a command such as "Do absolutely nothing until I call you!" a human would use common sense and, for example, attend to bodily functions in the meantime. A human would do *nothing about the relevant matter* until the event occurred. In addition, a human would recognize additional terminating events, such as a change in circumstances that make it impossible for the event to ever occur. A robot is likely to be constrained to a more literal interpretation, and unless it can infer a scope delimitation to the command, it would need to place the majority of its functions in abeyance. The faculties that would need to remain in operation are

- the sensory-perceptive subsystem needed to detect the condition,

- the recommencement triggering function,
- one or more daemons to provide a time-out mechanism (presumably the scope of the command is at least restricted to the expected remaining lifetime of the person who gave the command), and
- the ability to play back the audit trail so that an overseer can discover the condition on which the robot's resuscitation depends.

Asimov does not appear to have investigated whether the behavior of a robot in wait mode is affected by the laws. If it isn't, then it will not only fail to protect its own existence and to obey an order, but will also stand by and allow a human to be harmed. A robotic security guard could therefore be nullified by an attacker's simply putting it into a wait state.

Audit of robot compliance. For a fiction writer, it is sufficient to have the laws embedded in robots' positronic pathways (whatever they may be). To actually apply such a set of laws in robot design, however, it would be necessary to ensure that every robot

- had the laws imposed in precisely the manner intended, and
- was at all times subject to them — that is, they could not be overridden or modified.

It is important to know how malprogramming and modification of the laws' implementation in a robot (whether intentional or unintentional) can be prevented, detected, and dealt with.

In an early short story, robots were "rescuing" humans whose work required short periods of relatively harmless exposure to gamma radiation. Officials obtained robots with the first law modified so that they were incapable of injuring a human but under no compulsion to prevent one from coming to harm. This clearly undermined the remaining part of the first law, since, for example, a robot could drop a heavy weight toward a human, knowing that it would be fast enough and strong enough to catch it before it harmed the person. However, once gravity had taken over, the robot would be free to ignore the danger.[25] Thus, a partial implementation was shown to be risky, and the importance of robot audit underlined. Other risks include trapdoors, Trojan horses, and similar devices in the robot's programming.

A further imponderable is the effect of hostile environments and stress on the reliability and robustness of robots' performance in accordance with the laws. In one short story, it transpires that "The Machine That Won the War" had been receiving only limited and poor-quality data as a result of enemy action against its receptors and had been processing it unreliably because of a shortage of experienced maintenance staff. Each of the responsible managers had, in the interests of national morale, suppressed that information, even from one another, and had separately and independently "introduced a number of necessary biases" and "adjusted" the processing parameters in accordance with intuition. The executive director, even though unaware of the adjustments, had placed little reliance on the machine's output, preferring to carry out his responsibility to mankind by exercising his own judgment.[27]

A major issue in military applications generally[28] is the impossibility of contriving effective compliance tests for complex systems subject to hostile and competitive environments. Asimov points out that the difficulties of assuring compliance will be compounded by the design and manufacture of robots by other robots.[22]

Robot autonomy. Sometimes humans may delegate control to a robot and find themselves unable to regain it, at least in a particular context. One reason is that to avoid deadlock, a robot must be capable of making arbitrary decisions. Another is that the laws embody an explicit ability for a robot to disobey an instruction, by virtue of the overriding first law.

In an early Asimov short story, a robot "knows he can keep [the energy beam] more stable than we [humans] can, since he insists he's the superior being, so he must keep us out of the control room [in accordance with the first law]."[29] The same scenario forms the basis of one of the most vivid episodes in science fiction, HAL's attempt to wrest control of the spacecraft from Bowman in *2001: A Space Odyssey*. Robot autonomy is also reflected in a lighter moment in one of Asimov's later novels, when a character says to his companion, "For now I must leave you. The ship is coasting in for a landing, and I must stare intelligently at the computer that controls it, or no one will believe I am the captain."[14]

In extreme cases, robot behavior will involve subterfuge, as the machine determines that the human, for his or her own protection, must be tricked. In another early short story, the machines that manage Earth's economy implement a form of "artificial stupidity" by making intentional errors, thereby encouraging humans to believe that the robots are fallible and that humans still have a role to play.[23]

Scope for adaptation. The normal pattern of any technology is that successive generations show increased sophistication, and it seems inconceivable that robotic technology would quickly reach a plateau and require little further development. Thus there will always be many old models in existence, models that may have inherent technical weaknesses resulting in occasional malfunctions and hence infringement on the Laws of Robotics. Asimov's short stories emphasize that robots are leased from the manufacturer, never sold, so that old models can be withdrawn after a maximum of 25 years.

Looking at the first 50 years of software maintenance, it seems clear that successive modification of existing software to perform new or enhanced functions is one or more orders of magnitude harder than creating a new artifact to perform the same function. Doubts must exist about the ability of humans (or robots) to reliably adapt existing robots. The alternative — destruction of existing robots — will be resisted in accordance with the third law, robot self-preservation.

At a more abstract level, the laws are arguably incomplete because the frame of reference is explicitly human. No recognition is given to plants, animals, or as-yet-undiscovered (for example, extraterrestrial), intelligent life forms. Moreover, some future human cultures may place great value on inanimate creation, or on holism. If, however, late twentieth-century values have meanwhile been embedded in robots, that future culture may have difficulty wresting the right to change the values of the robots it has inherited. If machines are to have value sets, there must be a mechanism for adaptation, at least through human-imposed change. The difficulty is that most such value sets will be implicit rather than explicit; their effects will be scattered across a system rather than implemented in a modular and therefore replaceable manner.

At first sight, Asimov's laws are intuitively appealing, but their application encounters difficulties. Asimov, in his fiction, detected and investigated the laws' weaknesses, which this article (Part 1 of 2) has analyzed and classified. Part 2, in the next issue of *Computer*, will take the analysis further by considering the effects of Asimov's 1985 revision to the laws. It will then examine the extent to which the weaknesses in these laws may in fact be endemic to any set of laws regulating robotic behavior. ■

References

1. I. Asimov, *The Rest of the Robots* (a collection of short stories originally published between 1941 and 1957), Grafton Books, London, 1968.

2. N. Frude, *The Robot Heritage*, Century Publishing, London, 1984.

3. I. Asimov, *I, Robot* (a collection of short stories originally published between 1940 and 1950), Grafton Books, London, 1968.

4. I. Asimov, P.S. Warrick, and M.H. Greenberg, eds., *Machines That Think*, Holt, Rinehart, and Wilson, London, 1983.

5. I. Asimov, "Runaround" (originally published in 1942), reprinted in Reference 3, pp. 33-51.

6. L. Del Rey, "Though Dreamers Die" (originally published in 1944), reprinted in Reference 4, pp. 153-174.

7. I. Asimov, "Evidence" (originally published in 1946), reprinted in Reference 3, pp. 159-182.

8. H.M. Geduld and R. Gottesman, eds., *Robots, Robots, Robots*, New York Graphic Soc., Boston, 1978.

9. P.B. Scott, *The Robotics Revolution: The Complete Guide*, Blackwell, Oxford, 1984.

10. I. Asimov, *Robot Dreams* (a collection of short stories originally published between 1947 and 1986), Victor Gollancz, London, 1989.

11. A. Chandor, ed., *The Penguin Dictionary of Computers*, 3rd ed., Penguin, London, 1985.

12. I. Asimov, "The Bicentennial Man" (originally published in 1976), reprinted in Reference 4, pp. 519-561. Expanded into I. Asimov and R. Silverberg, *The Positronic Man*, Victor Gollancz, London, 1992.

13. A.C. Clarke and S. Kubrick, *2001: A Space Odyssey*, Grafton Books, London, 1968.

14. I. Asimov, *Robots and Empire*, Grafton Books, London, 1985.

15. I. Asimov, "Risk" (originally published in 1955), reprinted in Reference 1, pp. 122-155.

16. I. Asimov, *The Robots of Dawn*, Grafton Books, London, 1983.

17. I. Asimov, "Liar!" (originally published in 1941), reprinted in Reference 3, pp. 92-109.

18. I. Asimov, "That Thou Art Mindful of Him" (originally published in 1974), reprinted in *The Bicentennial Man*, Panther Books, London, 1978, pp. 79-107.

19. I. Asimov, *The Caves of Steel* (originally published in 1954), Grafton Books, London, 1958.

20. T. Winograd and F. Flores, *Understanding Computers and Cognition*, Ablex, Norwood, N.J., 1986.

21. I. Asimov, "Robbie" (originally published as "Strange Playfellow" in 1940), reprinted in Reference 3, pp. 13-32.

22. I. Asimov, *The Naked Sun* (originally published in 1957), Grafton Books, London, 1960.

23. I. Asimov, "The Evitable Conflict" (originally published in 1950), reprinted in Reference 3, pp. 183-206.

24. I. Asimov, "The Tercentenary Incident" (originally published in 1976), reprinted in *The Bicentennial Man*, Panther Books, London, 1978, pp. 229-247.

25. I. Asimov, "Little Lost Robot" (originally published in 1947), reprinted in Reference 3, pp. 110-136.

26. I. Asimov, "Robot Dreams," first published in Reference 10, pp. 51-58.

27. I. Asimov, "The Machine That Won the War" (originally published in 1961), reprinted in Reference 10, pp. 191-197.

28. D. Bellin and G. Chapman, eds., *Computers in Battle: Will They Work?* Harcourt Brace Jovanovich, Boston, 1987.

29. I. Asimov, "Reason" (originally published in 1941), reprinted in Reference 3, pp. 52-70.

Roger Clarke, reader in information systems at the Australian National University, has a background of 17 years in professional, managerial, and consulting roles within the information technology industry. Since 1988 he has directed a research program in supra-organizational systems, focusing on electronic commerce. His interests encompass organizational, economic, legal, and social aspects of information technology. He is also interested in the use of literature — particularly the anti-utopian and cyberpunk genres — as instruments of technological and social forecasting. Clarke has degrees from the University of New South Wales, Sydney. He is a member of the ACM and is active in the Australian Computer Society.

The author can be contacted at the Australian National University, Business Information Systems Group, Department of Commerce, GPO Box 4, Canberra, ACT 0200, Australia; Internet, roger.clarke@anu.edu.au.