COMPUTING MILIEUX

Asimov's Laws of Robotics: Implications for Information Technology

Part 2

Roger Clarke

Australian National University

Isaac Asimov's Laws of Robotics, first formulated in 1940, were primarily a literary device intended to support a series of stories about robot behavior. Over time, he found that the three laws included enough apparent inconsistencies, ambiguity, and uncertainty to provide the conflicts required for a great many stories. In examining the ramifications of these laws, Asimov revealed problems that might later confront real roboticists and information technologists attempting to establish rules for the behavior of intelligent machines.

With their fictional "positronic" brains imprinted with the mandate to (in order of priority) prevent harm to humans, obey their human masters, and protect themselves, Asimov's robots had to deal with great complexity. In a given situation, a robot might be unable to satisfy the demands of two equally powerful mandates and go into "mental freezeout." Semantics is also a problem. As demonstrated in Part 1 of this article (*Computer*, December 1993, pp. 53-61), language is much more than a set of literal meanings, and Asimov showed us that a machine trying to distinguish, for example, who or what is human may encounter many difficulties that humans themselves handle easily and intuitively. Thus, robots must have sufficient capabilities for judgment — capabilities that can cause them to frustrate the intentions of their masters when, in a robot's judgment, a higher order law applies.

As information technology evolves and machines begin to design and build other machines, the issue of human control gains greater significance. In time, human values tend to change; the rules reflecting these values, and embedded in existing robotic devices, may need to be modified. But if they are implicit rather than explicit, with their effects scattered widely across a system, they may not be easily replaceable. Asimov himself discovered many contradictions and eventually revised the Laws of Robotics.

Can a set of laws or rules reliably constrain the behavior of intelligent machines? Inadequacies in Asimov's laws suggest maybe not.

Asimov's 1985 revised Laws of Robotics

The zeroth law. After introducing the original three laws, Asimov detected, as early as 1950, a need to extend the first law, which protected individual humans, so that it would protect humanity as a whole. Thus, his calculating machines "have *the good of humanity* at heart through the overwhelming force of the First Law of Robotics"[1] (emphasis added). In 1985 he developed this idea further by postulating a "zeroth" law that placed humanity's interests above those of any individual while retaining a high value on individual human life.[2] The revised set of laws is shown in the sidebar.

Asimov pointed out that under a strict interpretation of the first law, a robot would protect a person even if the survival of humanity as a whole was placed at risk. Possible threats include annihilation by an alien or mutant human race, or by a deadly virus. Even when a robot's own powers of reasoning led it to conclude that mankind as a whole was doomed if it refused to act, it was nevertheless constrained: "I sense the oncoming of catastrophe . . . [but] I can only follow the Laws."[2]

Asimov's revised Laws of Robotics (1985)

Zeroth Law:
A robot may not injure humanity, or, through inaction, allow humanity to come to harm.

First Law:
A robot may not injure a human being, or, through inaction, allow a human being to come to harm, unless this would violate the Zeroth Law of Robotics.

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

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

In Asimov's fiction the robots are tested by circumstances and must seriously consider whether they can harm a human to save humanity. The turning point comes when the robots appreciate that the laws are indirectly modifiable by roboticists through the definitions programmed into each robot: "If the Laws of Robotics, even the First Law, are not absolutes, and if human beings can modify them, might it not be that perhaps, under proper conditions, we ourselves might mod — "[2] Although the robots are prevented by imminent "roblock" (robot block, or deadlock) from even completing the sentence, the groundwork has been laid.

Later, when a robot perceives a clear and urgent threat to mankind, it concludes, "Humanity as a whole is more important than a single human being. There is a law that is greater than the First Law: 'A robot may not injure humanity, or through inaction, allow humanity to come to harm.'"[2]

Defining "humanity." Modification of the laws, however, leads to additional considerations. Robots are increasingly required to deal with abstractions and philosophical issues. For example, the concept of humanity may be interpreted in different ways. It may refer to the set of individual human beings (a collective), or it may be a distinct concept (a generality, as in the notion of "the State"). Asimov invokes both ideas by referring to a tapestry (a generality) made up of individual contributions (a collective): "An individual life is one thread in the tapestry, and what is one thread compared to the whole? . . . Keep your mind fixed firmly on the tapestry and do not let the trailing off of a single thread affect you."[2]

A human roboticist raised a difficulty with the zeroth law immediately after the robot formulated it: "What is your 'humanity' but an abstraction? Can you point to humanity? You can injure or fail to injure a specific human being and understand the injury or lack of injury that has taken place. Can you see the injury to humanity? Can you understand it? Can you point to it?"[2] The robot later responds by positing an ability to "detect the hum of the mental activity of Earth's human population, overall. . . . And, extending that, can one not imagine that in the Galaxy generally there is the hum of the mental activity of all of humanity? How, then, is humanity an abstraction? It is something you can point to." Perhaps as Asimov's robots learn to reason with abstract concepts, they will inevitably become adept at sophistry and polemic.

The increased difficulty of judgment. One of Asimov's robot characters also points out the increasing complexity of the laws: "The First Law deals with specific individuals and certainties. Your Zeroth Law deals with vague groups and probabilities."[2] At this point, as he often does, Asimov resorts to poetic license and for the moment pretends that coping with harm to individuals does not involve probabilities. However, the key point is not affected: Estimating probabilities in relation to groups of humans is far more difficult than with individual humans.

> It is difficult enough, when one must choose quickly . . . , to decide which individual may suffer, or inflict, the greater harm. To choose between an individual and humanity, when you are not sure of what aspect of humanity you are dealing with, is so difficult that the very validity of Robotic Laws comes to be suspect. As soon as humanity in the abstract is introduced, the Laws of Robotics begin to merge with the Laws of Humanics — which may not even exist.[2]

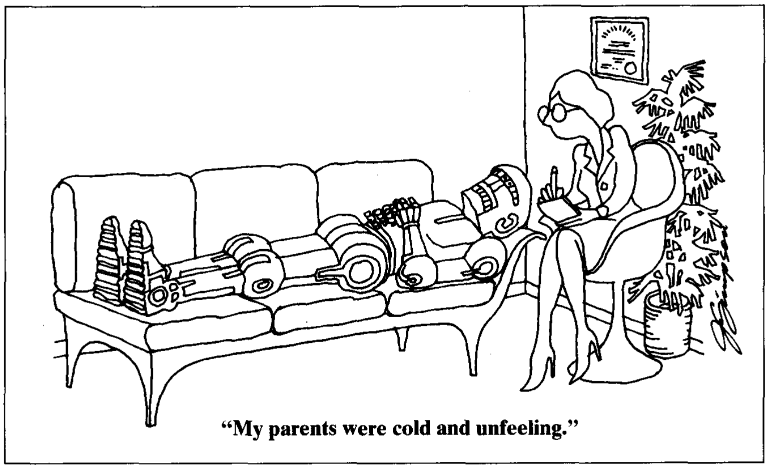
Robot paternalism. Despite these difficulties, the robots agree to implement the zeroth law, since they judge themselves more capable than anyone else of dealing with the problems. The original laws produced robots with considerable autonomy, albeit a qualified autonomy allowed by humans. But under the 1985 laws, robots were more likely to adopt a superordinate, paternalistic attitude toward humans.

Asimov suggested this when he first hinted at the zeroth law, because he had his chief robotpsychologist say that ". . . we can no longer understand our own creations. . . . [Robots] have progressed beyond the possibility of detailed human control."[1] In a more recent novella, a robot proposes to treat his form "as a canvas on which I intend to draw a man," but is told by the roboticist, "It's a puny ambition. . . . You're better than a man. You've gone downhill from the moment you opted for organicism."[3]

In the later novels, a robot with telepathic powers manipulates humans to act in a way that will solve problems,[4] although its powers are constrained by the psychological dangers of mind manipulation. Naturally, humans would be alarmed by the very idea of a mind-reading robot; therefore, under the zeroth and first laws, such a robot would be permitted to manipulate the minds of humans who learned of its abilities, making them forget the knowledge so that they could not be harmed by it. This is reminiscent of an Asimov story in which mankind is an experimental laboratory for higher beings[5] and Adams' altogether more flippant *Hitchhiker's Guide to the Galaxy*, in which the Earth is revealed as a large experiment in which humans are being used as laboratory animals by, of all things, white mice.[6] Someday those manipulators of humans might be robots.

Asimov's *The Robots of Dawn* is essentially about humans, with robots as important players. In the sequel *Robots and Empire*, however, the story is dominated by the two robots, and the humans seem more like their playthings. It comes as little surprise, then, that the robots eventually conclude that "it is not sufficient to be able to choose [among alternative humans or classes of human] . . . ; we must be able to shape."[2] Clearly, any subsequent novels in the series would have been about robots, with humans playing "bit" parts.

Robot dominance has a corollary that pervades the novels: History "grew less interesting as it went along; it became almost soporific."[4] With life's challenges removed, humanity naturally regresses into peace and quietude, becoming "placid, comfortable, and unmoving" — and stagnant.



"My parents were cold and unfeeling."

So who's in charge? As we have seen, the term human can be variously defined, thus significantly affecting the first law. The term humanity did not appear in the original laws, only in the zeroth law, which Asimov had formulated and enunciated by a robot.[2] Thus, the robots define human and humanity to refer to themselves as well as to humans, and ultimately to themselves alone. Another of the great science fiction stories, Clarke's *Rendezvous with Rama*,[7] also assumes that an alien civilization, much older than mankind, would consist of robots alone (although in this case Clarke envisioned biological robots). Asimov's vision of a robot takeover differs from those of previous authors only in that force would be unnecessary.

Asimov does *not* propose that the zeroth law must inevitably result in the ceding of species dominance by humans to robots. However, some concepts may be so central to humanness that any attempt to embody them in computer processing might undermine the ability of humanity to control its own fate. Weizenbaum argues this point more fully.[8]

The issues discussed here, and in Part 1, have grown increasingly speculative, and some are more readily associated with metaphysics than with contemporary applications of information technology. However, they demonstrate that even an intuitively attractive extension to the original laws could have very significant ramifications. Some of the weaknesses are probably inherent in any set of laws and hence in any robotic control regime.

Asimov's laws extended

The behavior of robots in Asimov's stories is not satisfactorily explained by the laws he enunciated. This section examines the design requirements necessary to effectively subject robotic behavior to the laws. In so doing, it becomes necessary to postulate several additional laws implicit in Asimov's fiction.

Perceptual and cognitive apparatus. Clearly, robot design must include sophisticated sensory capabilities. However, more than signal reception is needed. Many of the difficulties Asimov dramatized arose because robots were less than omniscient. Would humans, knowing that robots' cognitive capabilities are limited, be prepared to trust their judgment on life-and-death matters? For example, the fact that any single robot cannot harm a human does not protect humans from being injured or killed by robotic actions. In one story, a human tells a robot to add a chemical to a glass of milk and then tells another robot to serve the milk to a human. The result is murder by poisoning. Similarly, a robot untrained in first aid might move an accident victim and break the person's spinal cord. A human character in *The Naked Sun* is so incensed by these shortcomings that he accuses roboticists of perpetrating a fraud on mankind by omitting key words from the first law. In effect, it really means "A robot may do nothing that *to its knowledge* would injure a human being, and may not, through inaction, *knowingly* allow a human being to come to harm."[9]

Robotic architecture must be designed so that the laws can effectively control a robot's behavior. A robot requires a basic grammar and vocabulary to "understand" the laws and converse with humans. In one short story, a production accident results in a "mentally retarded" robot. This robot, defending itself against a feigned attack by a human, breaks its assailant's arm. This was not a breach of the first law, because it did not knowingly injure the human: "In brushing aside the threatening arm . . . it could not know the bone would break. In human terms, no moral blame can be attached to an individual who honestly cannot differentiate good and evil."[10] In Asimov's stories, instructions sometimes must be phrased carefully to be interpreted as mandatory. Thus, some authors have considered extensions to the apparatus of robots, for example, a "button labeled '*Implement Order*' on the robot's chest,"[11] analogous to the Enter key on a computer's keyboard.

A set of laws for robotics cannot be independent but must be conceived as part of a system. A robot must also be endowed with data collection, decision-analytical, and action processes by which it can apply the laws. Inadequate sensory, perceptual, or cognitive faculties would undermine the laws' effectiveness.

Without the First Law of Robotics. . .

"He was yelling at the robot, 'Am I the only one around here with a head on his shoulders? Now knock it off!'"

Additional implicit laws. In his first robot short story, Asimov stated that "long before enough can go wrong to alter that First Law, a robot would be completely inoperable. It's a mathematical impossibility [for Robbie the Robot to harm a human]."[12] For this to be true, robot design would have to incorporate a high-order controller (a "conscience"?) that would cause a robot to detect any potential for noncompliance with the laws and report the problem — or immobilize itself. The implementation of such a meta-law ("A robot may not act unless its actions are subject to the laws of robotics") might well strain both the technology and the underlying science. (Given the meta-language problem in twentieth-century philosophy, perhaps logic itself would be strained.) This difficulty highlights the simple fact that robotic behavior cannot be entirely automated; it is dependent on design and maintenance by an external agent.

Another of Asimov's requirements is that all robots must be subject to the laws at all times. Thus, it would have to be illegal for human manufacturers to create a robot that was not subject to the laws. In a future world that makes significant use of robots, their design and manufacture would naturally be undertaken by other robots. Therefore, the Laws of Robotics must include the stipulation that no robot may commit an act that could result in any robot's not being subject to the same laws.

The words "protect its own existence" raise a semantic difficulty. In *The Bicentennial Man*, Asimov has a robot achieve humanness by taking its own life. Van Vogt, however, wrote that "indoctrination against suicide" was considered a fundamental requirement.[13] The solution might be to interpret the word *protect* as applying to all threats, or to amend the wording to explicitly preclude self-inflicted harm.

Having to continually instruct robot slaves would be both inefficient and tiresome. Asimov hints at a further, deep-nested law that would compel robots to perform the tasks they were trained for:

> Quite aside from the Three Laws, there isn't a pathway in those brains that isn't carefully designed and fixed. We have robots planned for specific tasks, *implanted with specific capabilities*.[14] (Emphasis added.)

So perhaps we can extrapolate an additional, lower priority law: "A robot must perform the duties for which it has been programmed, except where that would conflict with a higher order law."

Asimov's laws regulate robots' transactions with humans and thus apply where robots have relatively little to do with one another or where there is only one robot. However, the laws fail to address the management of large numbers of robots. In several stories, a robot is assigned to oversee other robots. This would be possible only if each of the lesser robots were instructed by a human to obey the orders of its robot overseer. That would create a number of logical and practical difficulties, such as the scope of the human's order. It would seem more effective to incorporate in all subordinate robots an additional law, for example, "A robot must obey the orders given it by superordinate robots except where such orders would conflict with a higher order law." Such a law would fall between the second and third laws.

An extended set of the Laws of Robotics

The Meta-Law:
A robot may not act unless its actions are subject to the Laws of Robotics.

Law Zero:
A robot may not injure humanity, or, through inaction, allow humanity to come to harm.

Law One:
A robot may not injure a human being, or, through inaction, allow a human being to come to harm, unless this would violate a higher order law.

Law Two:
(a) A robot must obey orders given it by human beings, except where such orders would conflict with a higher order law.
(b) A robot must obey orders given it by superordinate robots, except where such orders would conflict with a higher order law.

Law Three:
(a) A robot must protect the existence of a superordinate robot as long as such protection does not conflict with a higher order law.
(b) A robot must protect its own existence as long as such protection does not conflict with a higher order law.

Law Four:
A robot must perform the duties for which it has been programmed, except where that would conflict with a higher order law.

The Procreation Law:
A robot may not take any part in the design, manufacture, or maintenance of a robot unless the new or modified robot's actions are subject to the Laws of Robotics.

Furthermore, subordinate robots should protect their superordinate robot. This could be implemented as an extension or corollary to the third law; that is, to protect itself, a robot would have to protect another robot on which it depends. Indeed, a subordinate robot may need to be capable of sacrificing itself to protect its robot overseer. Thus, an additional law superior to the third law but inferior to orders from either a human or a robot overseer seems appropriate: "A robot must protect the existence of a superordinate robot as long as such protection does not conflict with a higher order law."

The wording of such laws should allow for nesting, since robot overseers may report to higher level robots. It would also be necessary to determine the form of the superordinate relationships:

- a tree, in which each robot has precisely one immediate overseer, whether robot or human;
- a constrained network, in which each robot may have several overseers but restrictions determine who may act as an overseer; or
- an unconstrained network, in which each robot may have any number of other robots or persons as overseers.

This issue of a command structure is far from trivial, since it is central to democratic processes that no single entity shall have ultimate authority. Rather, the most senior entity in any decision-making hierarchy must be subject to review and override by some other entity, exemplified by the balance of power in the three branches of government and the authority of the ballot box. Successful, long-lived systems involve checks and balances in a lattice rather than a mere tree structure. Of course, the structures and processes of human organizations may prove inappropriate for robotic organization. In any case, additional laws of some kind would be essential to regulate relationships among robots.

The sidebar shows an extended set of laws, one that incorporates the additional laws postulated in this section. Even this set would not always ensure appropriate robotic behavior. However, it does reflect the implicit laws that emerge in Asimov's fiction while demonstrating that any realistic set of design principles would have to be considerably more complex than Asimov's 1940 or 1985 laws. This additional complexity would inevitably exacerbate the problems identified earlier in this article and create new ones.

While additional laws may be trivially simple to extract and formulate, the need for them serves as a warning. The 1940 laws' intuitive attractiveness and simplicity were progressively lost in complexity, legalisms, and semantic richness. Clearly then, formulating an actual set of laws as a basis for engineering design would result in similar difficulties and require a much more formal approach. Such laws would have to be based in ethics and human morality, not just in mathematics and engineering. Such a political process would probably result in a document couched in fuzzy generalities rather than constituting an operational-level, programmable specification.

Implications for information technologists

Many facets of Asimov's fiction are clearly inapplicable to real information technology or too far in the future to be relevant to contemporary applications. Some matters, however, deserve our consideration. For example, Asimov's fiction could help us assess the practicability of embedding some appropriate set of general laws into robotic designs. Alternatively, the substantive content of the laws could be used as a set of guidelines to be applied during the conception, design, development, testing, implementation, use, and maintenance of robotic systems. This section explores the second approach.

Recognition of stakeholder interests. The Laws of Robotics designate no particular class of humans (not even a robot's owner) as more deserving of protection or obedience than another. A human might establish such a relationship by command, but the laws give such a command no special status; another human could therefore countermand it. In short, the laws reflect the humanistic and egalitarian principles that theoretically underlie most democratic nations.

The laws therefore stand in stark contrast to our conventional notions about an information technology artifact, whose owner is implicitly assumed to be its primary beneficiary. An organization shapes an application's design and use for its own benefit. Admittedly, during the last decade users have been given greater consideration in terms of both the human-machine interface and participation in system development. But that trend has been justified by the better returns the organization can get from its information technology investment rather than by any recognition that users are stakeholders with a legitimate voice in decision making. The interests of other affected parties are even less likely to be reflected.

In this era of powerful information technology, professional bodies of information technologists need to consider

- identification of stakeholders and how they are affected;
- prior consultation with stakeholders;
- quality assurance standards for design, manufacture, use, and maintenance;
- liability for harm resulting from either malfunction or use in conformance with the designer's intentions; and
- complaint-handling and dispute-resolution procedures.

Once any resulting standards reach a degree of maturity, legislatures in the many hundreds of legal jurisdictions throughout the world would probably have to devise enforcement procedures.

The interests of people affected by modern information technology applications have been gaining recognition. For example, consumer representatives are now being involved in the statement of user requirements and the establishment of the regulatory environment for consumer electronic-funds-transfer systems. This participation may extend to the logical design of such systems. Other examples are trade-union negotiations with employers regarding technology-enforced change, and the publication of software quality-assurance standards.

For large-scale applications of information technology, governments have been called upon to apply procedures like those commonly used in major industrial and social projects. Thus, commitment might have to be deferred pending dissemination and public discussion of independent environmental or social impact statements. Although organizations that use information technology might see this as interventionism, decision making and approval for major information technology applications may nevertheless become more widely representative.

Closed-system versus open-system thinking. Computer-based systems no longer comprise independent machines each serving a single location. The marriage of computing with telecommunications has produced multicomponent systems designed to support all elements of a widely dispersed organization. Integration hasn't been simply geographic, however. The practice of information systems has matured since the early years when existing manual systems were automated largely without procedural change. Developers now seek payback via the rationalization of existing systems and varying degrees of integration among previously separate functions. With the advent of strategic and interorganizational systems, economies are being sought at the level of industry sectors, and functional integration increasingly occurs across corporate boundaries.

Although programmers can no longer regard the machine as an almost entirely closed system with tightly circumscribed sensory and motor capabilities, many habits of closed-system thinking remain. When systems have multiple components, linkages to other systems, and sophisticated sensory and motor capabilities, the scope needed for understanding and resolving problems is much broader than for a mere hardware/software machine. Human activities in particular must be perceived as part of the system. This applies to manual procedures within systems (such as reading dials on control panels), human activities on the fringes of systems (such as decision making based on computer-collated and -displayed information), and the security of the user's environment (automated teller machines, for example). The focus must broaden from mere technology to technology in use.

General systems thinking leads information technologists to recognize that relativity and change must be accommodated. Today, an artifact may be applied in multiple cultures where language, religion, laws, and customs differ. Over time, the original context may change. For example, models for a criminal justice system — one based on punishment and another based on redemption — may alternately dominate social thinking. Therefore, complex systems must be capable of adaptation.

Blind acceptance of technological and other imperatives. Contemporary utilitarian society seldom challenges the presumption that what *can* be done *should* be done. Although this technological imperative is less pervasive than people generally think, societies nevertheless tend to follow where their technological capabilities lead. Related tendencies include the economic imperative (what can be done more efficiently should be) and the marketing imperative (any effective demand should be met). An additional tendency might be called the "information imperative," the dominance of administrative efficiency, information richness, and rational decision making. However, the collection of personal data has become so pervasive that citizens and employees have begun to object.

The greater a technology's potential to promote change, the more carefully a society should consider the desirability of each application. Complementary measures that may be needed to ameliorate its negative effects should also be considered. This is a major theme of Asimov's stories, as he explores the hidden effects of technology. The potential impact of information technology is so great that it would be inexcusable for professionals to succumb blindly to the economic, marketing, information, technological, and other imperatives. Application software professionals can no longer treat the implications of information technology as someone else's problem but must consider them as part of the project.[15]

Human acceptance of robots. In Asimov's stories, humans develop affection for robots, particularly humaniform robots. In his very first short story, a little girl is too closely attached to Robbie the Robot for her parents' liking.[12] In another early story, a woman starved for affection from her husband and sensitively assisted by a humanoid robot to increase her self-confidence entertains thoughts approaching love toward it/him.[16]

Nonhumaniforms, such as conventional industrial robots and large, highly dispersed robotic systems (such as warehouse managers, ATMs, and EFT/POS systems) seem less likely to elicit such warmth. Yet several studies have found a surprising degree of identification by humans with computers.[17,18] Thus, some hitherto exclusively human characteristics are being associated with computer systems that don't even exhibit typical robotic capabilities.

If a robot-based economy develops without equitable adjustments, the backlash could be considerable.

Users must be continually reminded that the capabilities of hardware/software components are limited:

- They contain many inherent assumptions,
- they are not flexible enough to cope with all of the manifold exceptions that inevitably arise,
- they do not adapt to changes in their environment, and
- authority is not vested in hardware/software components but rather in the individuals who use them.

Educational institutions and staff training programs must identify these limitations; yet even this is not sufficient: The human-machine interface must reflect them. Systems must be designed so that users are required to continually exercise their own expertise, and system output should not be phrased in a way that implies unwarranted authority. These objectives challenge the conventional outlook of system designers.

Human opposition to robots. Robots are agents of change and therefore potentially upsetting to those with vested interests. Of all the machines so far invented or conceived of, robots represent the most direct challenge to humans. Vociferous and even violent campaigns against robotics should not be surprising. Beyond concerns of self-interest is the possibility that some humans could be revulsed by robots, particularly those with humanoid characteristics. Some opponents may be mollified as robotic behavior becomes more tactful. Another tenable argument is that by creating and deploying artifacts that are in some ways superior, humans degrade themselves.

System designers must anticipate a variety of negative reactions against their creations from different groups of stakeholders. Much will depend on the number and power of the people who feel threatened — and on the scope of the change they anticipate. If, as Asimov speculates,[9] a robot-based economy develops without equitable adjustments, the backlash could be considerable.

Such a rejection could involve powerful institutions as well as individuals. In one Asimov story, the US Department of Defense suppresses a project intended to produce the perfect robot-soldier. It reasons that the degree of discretion and autonomy needed for battlefield performance would tend to make robots rebellious in other circumstances (particularly during peace time) and unprepared to suffer their commanders' foolish decisions.[19] At a more basic level, product lines and markets might be threatened, and hence the profits and even the survival of corporations. Although even very powerful cartels might not be able to impede robotics for very long, its development could nevertheless be delayed or altered. Information technologists need to recognize the negative perceptions of various stakeholders and manage both system design and project politics accordingly.

The structuredness of decision making. For five decades there has been little doubt that computers hold significant computational advantages over humans. However, the merits of machine decision making remain in dispute. Some decision processes are highly structured and can be resolved using known algorithms operating on defined data items with defined interrelationships. Most structured decisions are candidates for automation, subject, of course, to economic constraints. The advantages of machines must also be balanced against risks. The choice to automate must be made carefully because the automated decision process (algorithm, problem description, problem-domain description, or analysis of empirical data) may later prove to be inappropriate for a particular type of decision. Also, humans involved as data providers, data communicators, or decision implementers may not perform rationally because of poor training, poor performance under pressure, or willfulness.

Unstructured decision making remains the preserve of humans for one or more of the following reasons:

- Humans have not yet worked out a suitable way to program (or teach) a machine how to make that class of decision.
- Some relevant data cannot be communicated to the machine.
- "Fuzzy" or "open-textured" concepts or constructs are involved.
- Such decisions involve judgments that system participants feel should not be made by machines on behalf of humans.

One important type of unstructured decision is problem diagnosis. As Asimov described the problem, "How . . . can we send a robot to find a flaw in a mechanism when we cannot possibly give precise orders, since we know nothing about the flaw ourselves? 'Find out what's wrong' is not an order you can give to a robot; only to a man."[20] Knowledge-based technology has since been applied to problem diagnosis, but Asimov's insight retains its validity: A problem may be linguistic rather than technical, requiring common sense, not domain knowledge. Elsewhere, Asimov calls robots "logical but not reasonable" and tells of household robots removing important evidence from a murder scene because a human did not think to order them to preserve it.[9]

The literature of decision support systems recognizes an intermediate case, semistructured decision making. Humans are assigned the decision task, and systems are designed to provide support for gathering and structuring potentially relevant data and for modeling and experimenting with alternative strategies. Through continual progress in science and technology, previously unstructured decisions are reduced to semistructured or structured decisions. The choice of which decisions to automate is therefore provisional, pending further advances in the relevant area of knowledge. Conversely, because of environmental or cultural change, structured decisions may not remain so. For example, a family of viruses might mutate so rapidly that the reference data within diagnostic support systems is outstripped and even the logic becomes dangerously inadequate.

A problem may be linguistic rather than technical, requiring common sense, not domain knowledge.

Delegating to a machine any kind of decision that is less than fully structured invites errors and mishaps. Of course, human decision-makers routinely make mistakes too. One reason for humans' retaining responsibility for unstructured decision making is rational: Appropriately educated and trained humans may make more right decisions and/or fewer seriously wrong decisions than a machine. Using common sense, humans can recognize when conventional approaches and criteria do not apply, and they can introduce conscious value judgments. Perhaps a more important reason is the arational preference of humans to submit to the judgments of their peers rather than of machines: If someone is going to make a mistake costly to me, better for it to be an understandably incompetent human like myself than a mysteriously incompetent machine.[8]

Because robot and human capabilities differ, for the foreseeable future at least, each will have specific comparative advantages. Information technologists must delineate the relationship between robots and people by applying the concept of decision structuredness to blend computer-based and human elements advantageously. The goal should be to achieve complementary intelligence rather than to continue pursuing the chimera of unneeded artificial intelligence. As Wyndham put it in 1932: "Surely man and machine are natural complements: They assist one another."[21]

Risk management. Whether or not subjected to intrinsic laws or design guidelines, robotics embodies risks to property as well as to humans. These risks must be managed; appropriate forms of risk avoidance and diminution need to be applied, and regimes for fallback, recovery, and retribution must be established.

Controls are needed to ensure that intrinsic laws, if any, are operational at all times and that guidelines for design, development, testing, use, and maintenance are applied. Second-order control mechanisms are needed to audit first-order control mechanisms. Furthermore, those bearing legal responsibility for harm arising from the use of robotics must be clearly identified. Courtroom litigation may determine the actual amount of liability, but assigning legal responsibilities in advance will ensure that participants take due care.

In most of Asimov's robot stories, robots are owned by the manufacturer even while in the possession of individual humans or corporations. Hence, legal responsibility for harm arising from robot noncompliance with the laws can be assigned with relative ease. In most real-world jurisdictions, however, there are enormous uncertainties, substantial gaps in protective coverage, high costs, and long delays.

Each jurisdiction, consistent with its own product liability philosophy, needs to determine who should bear the various risks. The law must be sufficiently clear so that debilitating legal battles do not leave injured parties without recourse or sap the industry of its energy. Information technologists need to communicate to legislators the importance of revising and extending the laws that assign liability for harm arising from the use of information technology.

Enhancements to codes of ethics. Associations of information technology professionals, such as the IEEE Computer Society, the Association for Computing Machinery, the British Computer Society, and the Australian Computer Society, are concerned with professional standards, and these standards almost always include a code of ethics. Such codes aren't intended so much to establish standards as to express standards that already exist informally. Nonetheless, they provide guidance concerning how professionals should perform their work, and there is significant literature in the area.

The issues raised in this article suggest that existing codes of ethics need to be reexamined in the light of developing technology. Codes generally fail to reflect the potential effects of computer-enhanced machines and the inadequacy of existing managerial, institutional, and legal processes for coping with inherent risks. Information technology professionals need to stimulate and inform debate on the issues. Along with robotics, many other technologies deserve consideration. Such an endeavor would mean reassessing professionalism in the light of fundamental works on ethical aspects of technology.

Asimov's Laws of Robotics have been a very successful literary device. Perhaps ironically, or perhaps because it was artistically appropriate, the sum of Asimov's stories disprove the contention that he began with: It is *not* possible to reliably constrain the behavior of robots by devising and applying a set of rules.

Tolerance and flexibility in design must replace the primacy of short-term objectives such as programming productivity.

The freedom of fiction enabled Asimov to project the laws into many future scenarios; in so doing, he uncovered issues that will probably arise someday in real-world situations. Many aspects of the laws discussed in this article are likely to be weaknesses in any robotic code of conduct. Contemporary applications of information technology such as CAD/CAM, EFT/POS, warehousing systems, and traffic control are already exhibiting robotic characteristics. The difficulties identified are therefore directly and immediately relevant to information technology professionals.

Increased complexity means new sources of risk, since each activity depends directly on the effective interaction of many artifacts. Complex systems are prone to component failures and malfunctions, and to intermodule inconsistencies and misunderstandings. Thus, new forms of backup, problem diagnosis, interim operation, and recovery are needed. Tolerance and flexibility in design must replace the primacy of short-term objectives such as programming productivity. If information technologists do not respond to the challenges posed by robotic systems, as investigated in Asimov's stories, information technology artifacts will be poorly suited for real-world applications. They may be used in ways not intended by their designers, or simply be rejected as incompatible with the individuals and organizations they were meant to serve. ■

References

1. I. Asimov, "The Evitable Conflict" (originally published in 1950), reprinted in I. Asimov, *I, Robot*, Grafton Books, London, 1968, pp. 183-206.

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

3. I. Asimov, "The Bicentennial Man" (originally published in 1976), reprinted in I. Asimov, P.S. Warrick, and M.H. Greenberg, eds., *Machines That Think*, Holt, Rinehart, and Wilson, 1983, pp. 519-561.

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

5. I. Asimov, "Jokester" (originally published in 1956), reprinted in I. Asimov, *Robot Dreams*, Victor Gollancz, London, 1989, pp. 278-294.

6. D. Adams, *The Hitchhiker's Guide to the Galaxy*, Harmony Books, New York, 1979.

7. A.C. Clarke, *Rendezvous with Rama*, Victor Gollancz, London, 1973.

8. J. Weizenbaum, *Computer Power and Human Reason*, W.H. Freeman, San Francisco, 1976.

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

10. I. Asimov, "Lenny" (originally published in 1958), reprinted in I. Asimov, *The Rest of the Robots*, Grafton Books, London, 1968, pp. 158-177.

11. H. Harrison, "War With the Robots" (originally published in 1962), reprinted in I. Asimov, P.S. Warrick, and M.H. Greenberg, eds., *Machines That Think*, Holt, Rinehart, and Wilson, 1983, pp. 357-379.

12. I. Asimov, "Robbie" (originally published as "Strange Playfellow" in 1940), reprinted in I. Asimov, *I, Robot*, Grafton Books, London, 1968, pp. 13-32.

13. A.E. Van Vogt, "Fulfillment" (originally published in 1951), reprinted in I. Asimov, P.S. Warrick, and M.H. Greenberg, eds., *Machines That Think*, Holt, Rinehart, and Wilson, 1983, pp. 175-205.

14. I. Asimov, "Feminine Intuition" (originally published in 1969), reprinted in I. Asimov, *The Bicentennial Man*, Panther Books, London, 1978, pp. 15-41.

15. R.A. Clarke, "Economic, Legal, and Social Implications of Information Technology," *MIS Quarterly*, Vol. 12, No. 4, Dec. 1988, pp. 517-519.

16. I. Asimov, "Satisfaction Guaranteed" (originally published in 1951), reprinted in I. Asimov, *The Rest of the Robots*, Grafton Books, London, 1968, pp. 102-120.

17. J. Weizenbaum, "Eliza," *Comm. ACM*, Vol. 9, No. 1, Jan. 1966, pp. 36-45.

18. S. Turkle, *The Second Self: Computers and the Human Spirit*, Simon & Schuster, New York, 1984.

19. A. Budrys, "First to Serve" (originally published in 1954), reprinted in I. Asimov, M.H. Greenberg, and C.G. Waugh, eds., *Robots*, Signet, New York, 1989, pp. 227-244.

20. I. Asimov, "Risk" (originally published in 1955), reprinted in I. Asimov, *The Rest of the Robots*, Grafton Books, London, 1968, pp. 122-155.

21. J. Wyndham, "The Lost Machine" (originally published in 1932), reprinted in A. Wells, ed., *The Best of John Wyndham*, Sphere Books, London, 1973, pp. 13-36, and in I. Asimov, P.S. Warrick, and M.H. Greenberg, eds., *Machines That Think*, Holt, Rinehart, and Wilson, 1983, pp. 29-49.

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, 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.