AP4CTE AP Seminar: Building a Dynamic Workforce
Research Strategies for Innovating and Problem-solving Across Career Paths
Module 5

Beyond Asimov: The Three Laws of Responsible Robotics

By Robin R. Murphy, *Texas A & M University*, and David D. Woods, *Ohio State University*

Published by the Institute for Human and Machine Cognition in *Human-Centered Computing*, July/August 2009

Since their codification in 1947 in the col-lection of short stories *I, Robot*, Isaac Asimov's three laws of robotics have been a staple of science fiction. Most of the stories assumed that the robot had complex perception and reasoning skills equivalent to a child and that robots were subservient to humans. Although the laws were simple and few, the stories attempted to demonstrate just how difficult they were to apply in various real-world situations. In most situations, although the robots usually behaved "logically," they often failed to do the "right" thing, typically because the particular context of application required subtle adjustments of judgment on the part of the robot (for example, determining which law took priority in a given situation, or what constituted helpful or harmful behavior).

The three laws have been so successfully inculcated into the public consciousness through entertainment that they now appear to shape society's expectations about how robots should act around humans. For instance, the media frequently refer to human–robot interaction in terms of the three laws. They've been the subject of serious blogs, events, and even scientific publications. The Singularity Institute organized an event and Web site, "Three Laws Unsafe," to try to counter public expectations of robots in the wake of the movie *I, Robot*. Both the philosophy[1] and AI[2] communities have discussed ethical considerations of robots in society using the three laws as a reference, with a recent discussion in *IEEE Intelligent Systems*.[3] Even medical doctors have considered robotic surgery in the context of the three laws.[4] With few notable exceptions,[5,6] there has been relatively little discussion of whether robots, now or in the near future, will have sufficient perceptual and reasoning capabilities to actually follow the laws. And there appears to be even less serious discussion as to whether the laws are actually viable as a framework for human–robot interaction, outside of cultural expectations.

Following the definitions in *Moral Machines: Teaching Robots Right from Wrong*,[7] Asimov's laws are based on functional morality, which assumes that robots have sufficient agency and cognition to make moral decisions. Unlike many of his successors, Asimov is less concerned with the details of robot design than in exploiting a clever literary device that lets him take advantage of the large gaps between aspiration and reality in robot autonomy. He uses the situations as a foil to explore issues such as

- the ambiguity and cultural dependence of language and behavior—for example, whether what appears to be cruel in the short run can actually become a kindness in the longer term;
- social utility—for instance, how different individuals' roles, capabilities, or backgrounds are valuable in different ways with respect to each other and to society; and
- the limits of technology—for example, the impossibility of assuring timely, correct actions in all situations and the omnipresence of trade-offs.

In short, in a variety of ways the stories test the lack of resilience in human–robot interactions.

The assumption of functional morality, while effective for entertaining storytelling, neglects operational morality. Operational morality links robot actions and inactions to the

decisions, assumptions, analyses, and investments of those who invent and make robotic systems and of those who commission, deploy, and handle robots in operational contexts. No matter how far the autonomy or robots ultimately advances, the important challenges of these accountability and liability linkages will remain.[8]

This essay reviews the three laws and briefly summarizes some of the practical shortcomings—and even dangers—of each law for framing human–robot relationships, including reminders about what robots can't do. We then propose an alternative, parallel set of laws based on what humans and robots can realistically accomplish in the foreseeable future as joint cognitive systems, and their mutual accountability for their actions from the perspectives of human-centered design and human– robot interaction.

Applying Asimov's Laws to Today's Robots

When we try to apply Asimov's laws to today's robots, we immediately run into problems. Just as for Asimov in his short stories, these problems arise from the complexities of situations where we would use robots, the limits of physical systems acting with limited resources in uncertain changing situations, and the interplay between the different social roles as different agents pursue multiple goals.

First Law

Asimov's first law of robotics states, "A robot may not injure a human being or, through inaction, allow a hu-man being to come to harm." This law is already an anachronism given the military's weaponization of robots, and discussions are now shifting to the question of whether weaponized robots can be "humane."[9,10] Such weaponization is no longer limited to situations in which humans remain in the loop for control. The South Korean government has published videos on YouTube of robotic border-security guards. Scenarios have been proposed where it would be permissible for a military robot to fire upon anything moving (presumably targeting humans) without direct human permission.[11]

Even if current events hadn't made the law irrelevant, it's moot because robots cannot infallibly recognize humans, perceive their intent, or reliably interpret contextualized scenes. A quick review of the computer vision literature shows that scientists continue to struggle with many fundamental perceptual processes. Current commercial security packages for recognizing the face of a person standing in a fixed position continue to fall short of expectations in practice. Many robots that "recognize" humans use indirect cues such as heat and motion, which only work in constrained contexts. These problems confirm Norbert Wiener's warnings about such failure possibilities.[8] Just as he envisioned many years ago, today's robots are literal-minded agents—that is, they can't tell if their world model is the world they're really in.

All this aside, the biggest problem with the first law is that it views safety only in terms of the robot—that is, the robot is the responsible safety agent in all matters of human–robot interaction. While some speculate on what it would mean for a robot to be able to discharge this responsibility, there are serious practical, theoretical, social-cognitive, and legal limitations.[8,12] For example, from a legal perspective the robot is a product, so it's not the responsible agent. Rather, the robot's owner or manufacturer is liable for its actions. Unless robots are granted a person-equivalent status, somewhat like corporations are now legally recognized as individual entities, it's difficult to imagine standard product liability law not applying to them. When a failure occurs, violating Asimov's first law, the human stakeholders affected by that failure will

engage in the processes of causal attribution. Afterwards, they'll see the robot as a device and will look for the person or group who set up or instructed the device erroneously or who failed to supervise (that is, stop) the robot before harm occurred. It's still commonplace after accidents for manufacturers and organizations to claim the result was due only to human error, even when the system in question was operating autonomously.[8,13]

Accountability is bound up with the way we maintain our social relationships. Human decision-making always occurs in a context of expectations that one might be called to account for his or her decisions. Expectations for what's considered an adequate explanation and the consequences for people when their explanation is judged inadequate are critical parts of accountability systems—a reciprocating cycle of being prepared to provide an accounting for one's actions and being called by others to provide an account. To be considered moral agents, robots would have to be capable of participating personally in this reciprocating cycle of accountability—an issue that, of course, concerns more than any single agent's capabilities in isolation.

Second Law

Asimov's second law of robotics states, "A robot must obey orders given to it by human beings, except where such orders would conflict with the first law." Although the law itself takes no stand on how humans would give orders, Asimov's robots relied on their understanding of verbal directives. Unfortunately, robust natural language understanding still continues to lie just beyond the frontiers of today's AI.[14] It's true that, after decades of research, computers can now construct words from phonemes with some consistency—as improvements in voice dictation and call centers attest. Language-understanding capabilities also work well for specific types of well-structured tasks. However, the goal of meaningful machine participation in open-ended conversational con-texts remains elusive. Additionally, we must account for the fact that not all directives are given verbally. Humans use gestures and add affect through body posture, facial expressions, and motions for clarification and emphasis. Indeed, high-performance, experienced teams use highly pointed and coded forms of verbal and nonverbal communication in fluid, interdependent, and idiosyncratic ways.

What's more interesting about the second law from a human–robot interaction standpoint is that at its core, it almost captures the more important idea that intelligent robots should notice and take stock of humans (and that the people robots encounter or interact with can notice pertinent aspects of robots' behavior).[15] For example, it is acceptable for a robot to merely not hit a person in a hospital hall, or should it conform to social convention and acknowledge the person in some way ("excuse me" or a nod of a camera pan-tilt)? Or if a robot operating in public places included two-way communication devices, could a bystander recognize that the robot provided a means to report a crime or a fire?

Third Law

The third law states, "A robot must protect its own existence as long as such protection does not conflict with the first or second law." Because today's robots are expensive, you'd think designers would be naturally motivated to incorporate some form of the third law into their products. For example, even the inexpensive iRobot Roomba detects stairs that could cause a fatal fall. Surprisingly, however, many expensive commercial robots lack the means to fully protect their owners' investment.

An extreme example of this is in the design of robots for military applications or bomb squads. Such robots are designed to be teleoperated by a person who bears full responsibility for all safety matters. Human-factors studies show that remote operators are immediately at a disadvantage, working through a mediated interface with a time delay. Worse yet, remote operators are required to operate the robot through poor human–computer interfaces and in contexts where the operator can be fatigued, overloaded, or under high stress. As a result, when an abnormal event occurs, they may be distracted or not fully engaged and thus might not respond adequately in time. The result for a robot is akin to expecting an astronaut on a planet's surface to request and wait for permission from mission control to perform even simple reflexes such as ducking.

What is puzzling about today's limited attempts to conform to the third law is that there are well-established technological solutions for basic robot survival activities that work for autonomous and human-controlled robots. For instance, since the 1960s we've had technology to assure guarded motion, where the human drives the robot but onboard software will not allow the robot to make potentially dangerous moves (for example, collide with obstacles or exceed speed limits or boundaries) without explicit orders (an implicit invocation of the second law). By the late 1980s, guarded motion was encapsulated into tactical reactive behaviors, essentially giving robots reflexes and tactical authority. Perhaps the most important reason that guarded motion and reflexive behaviors haven't been more widely deployed is that they require additional sensors, which would add to the cost. This increase in cost may not appear to be justified to customers, who tend to be wildly overconfident that trouble and complexities outside the bounds of expected behavior rarely arise.

The Alternative Three Laws of Responsible Robotics

To address the difficulties of applying Asimov's three laws to the current generation of robots while respecting the laws' general intent, we suggest the three laws of responsible robotics.

Alternative First Law

Our alternative to Asimov's first law is "A human may not deploy a robot without the human–robot work system meeting the highest legal and professional standards of safety and ethics." Since robots are indeed subject to safety regulations and liability laws, the requirement of meeting legal standards for safety would seem self-evident. For instance, the medical-device community has done extensive research to validate robot sensing of scalpel pressures and tissue contact parameters, and it invests in failure mode and effect analyses (consistent with FDA medical-device standards).

In contrast, mobile roboticists have a somewhat infamous history of disregarding regulations. For example, robot cars operating on public roads, such as those used in the DARPA Urban Grand Challenge, are considered by US federal and state transportation regulations as experimental vehicles. Deploying such vehicles requires voluminous and tedious permission applications. Regrettably, the 1995 CMU "No Hands Across America" team neglected to get all appropriate permissions while driving autonomously from Pittsburgh to Los Angeles, and were stopped in Kansas. The US Federal Aviation Ad-ministration makes a clear distinction between flying unmanned aerial vehicles as a hobby and flying them for R&D or commercial practices, effectively slowing or stopping many R&D efforts. In response to these difficulties, a culture

preferring "forgiveness" to "permission" has grown up in some research groups. Such attitudes indicate a poor safety culture at universities that could, in turn, propagate to government or industry. On the positive side, the robot competitions sponsored by the Association for Unmanned Vehicle Systems International are noteworthy in their insistence on having safe areas of operation, clear emergency plans, and safety officers present.

Meting the minimal legal requirements is not enough—the alternative first law demands the highest professional ethics in robot deployment. A failure or accident involving a robot can effectively end an entire branch of robotics research, even if the operators aren't legally culpable. Responsible communities should proactively consider safety in the broadest sense, and funding agencies should find ways to increase the priority and scope of research funding specifically aimed at relevant legal concerns.

The highest professional ethics should also be applied in product development and testing. Autonomous robots have known vulnerabilities to problems stemming from interrupted wireless communications. Signal reception is impossible to predict, yet robust "return to home if signal lost" and "stop movement if GPS lost" functionality hasn't yet become an expected component of built-in robot behavior. This means robots are operating counter to reasonable and prudent assumptions. Worse yet, when they're operating experimentally, robots often encounter unanticipated factors that affect their control. Simply saying an unfortunate event was unpredictable doesn't relieve the designers of responsibility. Even if a specific disturbance is unpredictable in detail, the fact that there will be disturbances is virtually guaranteed, and designing for resilience in the face of these is fundamental.

As a matter of professional common sense, robot design should start with safety first, then add the interesting software and hardware. Ro-bots should carry "black boxes" or recorders to show what they were doing when a disturbance occurred, not only for the sake of an accident investigation but also to trace the robots' behavior in context to aid diagnosis and debugging. There should be a formal safety plan and checklists for contingencies. These do not have to be extensive and time consuming to be effective. A litmus test for developers might be "If a group of experts from the IEEE were to write about your robot after an accident, what would they say about system safety and your professionalism?" Fundamentally, the alternative first law places responsibility for safety and efficacy on humans within the larger social and environmental context in which robots are developed, deployed, and operated.

Alternative Second Law

As an alternative to Asimov's second law, we propose the following: "A robot must respond to humans as appropriate for their roles." The capability to respond appropriately—*responsiveness*—may be more important to human–robot interaction than the capability of autonomy. Not all robots will be fully autonomous over all conditions. For example, a robot might be constrained to follow waypoints but will be expected to generate appropriate responses to people it encounters along the way. Responsiveness depends on the social environment, the kinds of people and their expectations that a robot might encounter in its work envelope. Rather than assume the relationship is hierarchical with the human as the superior and the robot as the subordinate so that all communication is a type of order, the alternative second law states that robots must be built so that the interaction fits the relationships and roles of each member in a given environment. The relationship determines the degree to which a robot is obligated to respond. It might ignore a hacker completely. Orders exceeding the authority of the speaker

might be disposed of politely ("please have your superior confirm your request") or with a warning ("interference with a law enforcement robot may be a violation"). Note that defining "appropriate response" may address concerns about robots being abused.[16]

The relationship also determines the mode of the response. How the robot signals or expresses itself should be consistent with that relationship. Casual relationships might rely on natural language, whereas trained teams performing specific tasks could coordinate activities through other signals such as body position and gestures.

The requirement for responsiveness captures a new form of autonomy (not as isolated action but the more difficult behavior of engaging appropriately with others). However, developing robots' capability for responsiveness requires a significant research effort, particularly in how robots can perceive and identify the different members, roles, and cues of social environment.

Alternative Third Law

Our third law is "A robot must be endowed with sufficient situated autonomy to protect its own existence as long as such protection provides smooth transfer of control to other agents consistent with the first and second laws." This law specifies that a human–robot system should be able to transition smoothly from whatever degree of autonomy or roles the robots and humans were inhabiting to a new control relationship given the nature of the disruption, impasse, or opportunity encountered or anticipated. When developers focus narrowly on the goal of isolated autonomy and fall prey to overconfidence by underestimating the potential for surprises to occur, they tend to minimize the importance of transfer of control. But bumpy transfers of control have been noted as a basic difficulty in human interaction with automation that can contribute to failures.[17]

The alternative third law addresses situated autonomy and smooth transfer of control, both of which interact with the prescriptions of the other laws. To be consistent with the second law requires that humans in a given role might not always have complete control of the robot (for example, when conditions require very short reaction times, a pilot may not be allowed to override some commands generated by algorithms that attempt to provide envelope protection for the aircraft). This in turn implies that an aspect of the design of roles is the identification of classes of situations that demand transfer of control, so that the exchange processes can be specified as part of roles. This is when the human takes control from the robot for a specialized aspect of the mission in anticipation of conditions that will challenge the limits of the robot's capabilities, or in an emergency. De-cades of human factors research on human out-of-the-loop control problems, handling of anomalies, cascades of disturbances, situation awareness, and autopilot/pilot transfer of control can inform such designs.

To be consistent with the first law requires designers to explicitly address what is the appropriate situated autonomy (for example, identifying when the robot is better informed or more capable than the human owing to latency, sensing, and so on) and to provide mechanisms that permit smooth transfer of control. To disregard the large body of literature on resilience and failure due to bumpy transfer of control would violate the designers' ethical obligation.
The alternative second and third laws encourage some forms of increased autonomy related to responsiveness and the ability to engage in various forms of smooth transfer of control. To be able to engage in these activities with people in various roles, the robot will need more situated intelligence. The result is an irony that has been noted before: increased capability for autonomy

and authority leads to the need to participate in more sophisticated forms of coordinated activity.[8]

Table 1. Asimov's laws of robotics versus the alternative laws of responsible robotics

	Asimov's laws	Alternative laws
1	A robot may not injure a human being or, through inaction, allow a human being to come to harm.	A human may not deploy a robot without the human–robot work system meeting the highest legal and professional standards of safety and ethics.
2	A robot must obey orders given to it by human beings, except where such orders would conflict with the first law.	A robot must respond to humans as appropriate for their roles.
3	A robot must protect its own existence as long as such protection does not conflict with the first or second law.	A robot must be endowed with sufficient situated autonomy to protect its own existence as long as such protection provides smooth transfer of control to other agents consistent the first and second laws.

Discussion

Our critique reveals that robots need two key capabilities: responsiveness and smooth transfer of control. Our proposed alternative laws remind robotics researchers and developers of their legal and professional responsibilities. They suggest how people can conduct human–robot interaction re-search safely, and they identify critical research questions.

Table 1 places Asimov's three laws side by side with our three alternative laws. Asimov's laws assume functional morality—that robots are capable of making (or are permitted to make) their own decisions—and ignore the legal and professional responsibility of those who design and deploy them (operational morality). More importantly for human–robot interaction, Asimov's laws ignore the complexity and dynamics of relationships and responsibilities between robots and people and how those relationships are expressed. In contrast, the alternative three laws emphasize responsibility and resilience, starting with enlightened, safety-oriented designs (alter-native first law), then adding responsiveness (alternative second law) and smooth transfer of control (alternative third law).

The alternative laws are designed to be more feasible to implement than Asimov's laws given current technology, although they also raise critical questions for research. For example, the alternative first law isn't concerned with technology per se but with the need for robot developers to be aware of human systems design principles and to take responsibility proactively for the consequences of errors and failures in human–robot systems. Standard tools from the aerospace, medical, and chemical manufacturing safety cultures, including training, formal processes, checklists, black boxes, and safety officers, can be adopted. Network and physical security should be incorporated into robots, even during development.

The alternative second and third laws require new research directions for robotics to leverage and build on existing results in social cognition, cognitive engineering, and resilience engineering. The laws suggest that the ability for robots to express relationships and obligations through social roles will be essential to all human–robot interaction. For example, work on entertainment robots and social robots provides insights about how robots can express emotions or affect appropriate to people they encounter. The extensive literature from cognitive engineering on transfer of control and general human out-of-the-loop control problems can be redirected at robotic systems. The techniques for resilience engineering are beginning to identify new control architectures for distributed, multi-echelon systems that include systems that include robots.

The fundamental difference between Asimov's laws, which focus on robots' functional morality and full moral agency, and the alternative laws, which focus on system responsibility and resilience, illustrates why the robotics community should resist public pressure to frame current human–robot interaction in terms of Asimov's laws. Asimov's laws distract from capturing the diversity of robotic missions and initiative. Understanding these diversities and complexities is critical for designing the "right" interaction scheme for a given domain.

Ironically, Asimov's laws really are robot-centric because most of the initiative for safety and efficacy lies in the robot as an autonomous agent. The alternative laws are human-centered because they take a systems approach. They emphasize that

- responsibility for the consequences of robots' successes and failures lies in the human groups that have a stake in the robots' activities, and
- capable robotic agents still exist in a web of dynamic social and cognitive relationships.

Ironically, meeting the requirements of the alternative laws leads to the need for robots to be more capable agents—that is, more responsive to others and better at interaction with others.

We propose the alternative laws as a way to stimulate debate about robots' accountability when their actions can harm people or human interests. We also hope that these laws can serve to direct R&D to enhance human–robot systems. Finally, while perhaps not as entertaining as Asimov's laws, we hope the alternative laws of responsible robotics can better communicate to the general public the complex mix of opportunities and challenges of robots in today's world.

Acknowledgments

We thank Jeff Bradshaw, Cindy Bethel, Jenny Burke, Victoria Groom, and Leila Takayama for their helpful feedback and Sung Huh for additional references. The second author's contributions was based on participation in the Advanced Decision Architectures Collaborative Technology Alliance, sponsored by the Us Army Research Laboratory under cooperative agreement DAAD190-01-2-0009.

References

1. S.L. Anderson, "Asimov's 'Three Laws of Robotics' and Machine Metaethics," *AI and Society*, vol. 22, no. 4, 2008, pp. 477–493.
2. A. Sloman, "Why Asimov's Three Laws of Robotics are Unethical," 27 July 2006; www.cs.bham.ac.uk/research/projects/cogaff/misc/asimov-three-laws. html.
3. C. Allen, W. Wallach, and I. Smit, "Why Machine Ethics*?" IEEE Intelligent Systems*, vol. 21, no. 4, 2006, pp. 12–17.
4. M. Moran, "Three Laws of Robotics and Surgery," J. *Endourology*, vol. 22, no. 8, 2008, pp. 1557–1560.
5. R. Clarke, "Asimov's Laws of Robotics: Implications for Information Technology Part 1," *Computer*, vol. 26, no. 12, 1993, pp. 53–61.
6. R. Clarke, "Asimov's Laws of Robotics: Implications for Information Technology Part 2," *Computer*, vol. 27, no. 1, 1994, pp. 57–66.
7. W. Wallach and C. Allen, *Moral Machines: Teaching Robots Right from Wrong*, Oxford Univ. Press, 2009.
8. D. Woods and E. Hollnagel, *Joint Cognitive Systems: Patterns in Cognitive Systems Engineering*, Taylor and Francis, 2006.
9. R.C. Arkin and L. Moshkina, "Lethality and Autonomous Robots: An Ethical Stance," *Proc. IEEE Int'l Symp. Technology and Society* (ISTAS 07), IEEE Press, 2007, pp. 1–3.
10. N. Sharkey, "The Ethical Frontiers of Robotics," *Science*, vol. 322, no. 5909, 2008, pp. 1800–1801.
11. M.F. Rose et al., *Technology Development for Army Unmanned Ground Vehicles*, Nat'l Academy Press, 2002.

12. D. Woods, "Conflicts between Learning and Accountability in Patient Safety," *DePaul Law Rev*., vol. 54, 2005, pp. 485–502.
13. S.W.A. Dekker, *Just Culture: Balancing Safety and Accountability*, Ashgate, 2008.
14. J. Allen et al., "Towards Conversational Human–Computer Interaction," *AI Magazine*, vol. 22, no. 4, 2001, pp. 27–38.
15. J.M. Bradshaw et al., "Dimensions of Adjustable Autonomy and Mixed-Initiative Interaction," *Agents and Computational Autonomy: Potential, Risks, and Solutions*, M. Nickles, M. Rovatsos, and G. Weiss, eds., LNCS 2969, Springer, 2004, pp. 17–39.
16. B. Whitby, "Sometimes It's Hard to Be a Robot: A Call for Action on the Ethics of Abusing Artificial Agents," *Interacting with Computers*, vol. 20, no. 3, 2008, pp. 326–333.
17. D.D. Woods and N. Sarter, "Learning from Automation Surprises and 'Going Sour' Accidents," *Cognitive Engineering in the Aviation Domain*, N.B. Sarter and R. Amalberti, eds., Nat'l Aeronautics and Space Administration, 1998.

Robin Murphy is the Raytheon Professor of Computer Science and Engineer-ing at Texas A&M University. Contact him at murphy@cse.tamu.edu.

David D. Woods is a professor in the Human Systems Integration section of the Department of Integrated Systems Engineering at Ohio State University. Contact him at woods.2@osu.edu.