Can Robots Be Trusted?

Humans have feared a robotic uprising since the machines first appeared in science fiction. Today, experts caution against a more insidious threat. We might like living with them too much.

by erik sofge
photographs by gregg segal

Being hacked by a robot requires much less hardware than I expected. There's no need for virtual-reality goggles or 3D holograms. There are no skullcaps studded with electrodes, no bulky cables or hair-thin nanowires snaking into my brain. Here's what it takes: one pair of alert, blinking eyeballs.

I'm in the Media Lab, part of MIT's sprawling campus in Cambridge, Mass. Like most designated research areas, the one belonging to the Personal Robots Group looks more like a teenage boy's bedroom than some pristine laboratory—it bursts with knotted cables, old pizza boxes and what are either dissected toys or autopsied robots. Amid the clutter, a 5-foot-tall, three-wheeled humanoid robot boots up and starts looking around the room. It's really looking, the oversize blue eyes tracking first, and the white, swollen, doll-like head following, moving and stopping as though focusing on each researcher's face. Next turns, looks at me. The eyes blink. I stop talking, mid-sentence, and look back. It's as instinctive as meeting a newborn's roving eyes. What do you want? I feel like asking. What do you need? If I was hoping for dispassionate, journalistic distance—and I was—I never had a chance.

"Right now it's doing a really basic look-around," researcher Matt Berlin says. "I think it's happy, because it has a face to look at." In another kind of robotics lab, a
Capable of gesturing and speaking, Sarcos was built in 1997 to talk about technology with children at the Carnegie Science Center in Pittsburgh. Social robots like it are expected to be a $15 billion industry by 2015.
humanoid bot might be motivated by a specific physical goal—
cross the room without falling, find the appropriate colored 
ball and give it a swift little kick. Nexi’s functionality is more 
ineffable. This is a social robot. Its sole purpose is to interact 
with people. Its mission is to be accepted.

That’s a mission any truly self-aware robot would probably 
turn down. To gain widespread acceptance could mean fighting 
decades of robot-related fear and loathing. Such stigmas 
rang from doomsday predictions of machines that inevitably 
will wage war on mankind to the belief that humanoid robots will 
always be hopelessly unerving and unsuitable companions.

For Nexi, arguably the biggest star of the human–robot inter-
action (HRI) research field, fame is already synonymous with 
fear. Before visiting the Media Lab, I watched a video of Nexi 
that’s been seen by thousands of people on YouTube. Nexi rolls 
into view, pivots stiffly to face the camera and introduces itself 
in a perfectly pleasant female voice. If the goal was to make Nexi 
early, the clip is a disaster. The eyes are big and expressive, 
the face is childish and cute, but everything is just slightly off, 
like a possessed doll masquerading as a giant toddler. Or, for 
the existentially minded, something more deeply disturbing— 
a robot with real emotions, equally capable of loving and despising 
you. Viewers dubbed its performance “creepy.”

Now, staring back at Nexi, I’m an instant robot apologist. 
I want to shower those clips with embarrassingly positive com-
ments, to tell the haters and the doubters that the future of HRI 
is bright. There’s no way seniors will reject the medias handed 
by their children, winking live-in-nurse bots. Children, no 
doubt, will love day-care robots, even if the bots sometimes fail 
to console them, or grind to an unresponsive halt because of 
buggy software or faulty battery packs. To turn today’s faceless 
Roboas into tomorrow’s active, autonomous machine compan-
iens, social robots need only to follow Nexi’s example, tapping 
into powerful, even uncontrollable human instincts.

That’s why Nexi’s metallic arms and hands are drifting 
around in small, lifelike movements. It’s why Nexi searches for 
faces and seems to look you in the eye. When it blinks 
again, with a little motorized buzz, I realize I’m smiling at this thing. I’m responding to it 
as one social, living creature to another. 
Nexi hasn’t said a word, and I already want to be its friend.

As it turns out, knowing your brain is being hacked by a robot 
doesn’t make it any easier to resist. And perhaps that’s the real danger 
of social robots. While humans have been busy hypothesizing about malevolent computers and the limits of rubber flesh, roboticists may have stumbled onto a more 
genuine threat. When face to face with actual robots, people may become too 
attached. And like human relationships, those attachments can be fraught with pitfalls: How will grandma feel, for example, when her companion bot is packed off for an 
upgrade and comes back a complete stranger?

When a machine can push our Darwinian buttons so easily, 
discharging our deep-seated reservations with a well-timed flut-
ter of its artificial eyelids, maybe fear isn’t such a stupid reaction 
after all. Maybe we’ve just been afraid of the wrong thing.

Robots began scaring us long before they existed. In 1921, the 
Czech play R.U.R., or Rossum’s Universal Robots, simultane-
ously introduced the word “robot” and the threat of a robot 
apocalypse. In a proclamation issued in the play’s first act, the 
robots, built as cheap, disposable laborers, make their intentions 
clear: “Robots of the world, we enjoin you to exterminate 
mankind. Don’t spare the men. Don’t spare the women.” The 
origins of the evil robot can be traced back even further (see 
page 59), but R.U.R.’s new species of hoxman was all the rage 
in the pulp sci-fi of the ’40s and ’50s—well before the actual 
research field of robotics. In fact, I, Robot author Isaac Asimov 
coined the term “robotics” at the same time that he began 
developing ethical laws for robots in his short stories.

By the time Arnold Schwarzenegger’s T-800 gunned down 
an entire police precinct in the 1984 movie The Terminator, the 
robot insurgency had become one of pop culture’s most 
entrenched clichés. The film has since become shorthand for 
a specific fear: that artificial intelligence (AI) will become too 
telligent, too obsessed with self-preservation. The Terminator 
colors the way we think about robots, AI and even the booming 
business of unmanned warfare. The Office of Naval Research, 
among others, has studied whether ethical guidelines will be 
needed for military robots, and in a 2008 preliminary report 
the authors tackle the bleakest possible endgame: “Terminator 
scenarios where machines turn against us lesser humans.”

But according to Patrick Lin, an assistant professor of phil-
osophy at California Polytechnic State University and an ethics 
Fellow at the U.S. Naval Academy, the need for ethical bots isn’t 
restricted to the battlefield. “Social robots probably pose a 
greater risk to the average person than a military robot,” Lin 
says. “They won’t be armed, but we will be coming face to face 
with them, quite soon.”

That, of course, is precisely the kind of quote reporters work hard to publish. The media 
homes in on juicy details about the hypoth-
etical danger of self-organizing AI, and the prospect of amoral robots 
gunning down civilians. But the real 
threats posed by robots may have 
nothing to do with the Terminator 
scenario. Because compared to 
even the dumbest armed insurgent, 
robots are practically brain-dead.

Take Nexi, for example. Considered 
to be one of the most advanced 
social robots in the world, Nexi can 
understand only the most basic vocal 
instructions. During my visit, it couldn’t 
even do that—it was in the process of being 
loaded with behavioral software developed 
for another MIT robot, the fuzzy, 
big-eared Leonardo. Now in semi-retirement—its 
motors have gone rickety—Leonardo learns
Social robots probably pose a greater risk to the average person than a military robot," Lin says. "They won't be armed, but we will be coming face to face with them, quite soon."
from humans such lessons as which blocks fit into a given puzzle, or which stuffed animal is “good” and which it should be afraid of. The implications are of the mind-blowing variety: a robot that listens to what we say and learns to crave or fear what we tell it to. Programmed with Leonardo’s smarts, “maybe in a year Nexi will be able to have a conversation with you that’s very boring,” MIT’s Berlin says. “But it may be pretty interesting if you’re trying to escape a burning building.”

If David Hanson, the founder of Hanson Robotics, has his way, the Texas-based company’s latest social robot, Zeno, could be talking circles around Nexi by the end of this year. At $2500, the 23-inch-tall humanoid robot would be a bargain, not because of its hardware but because of the code crammed into its cartoonish head. “The intelligent software can be aware of multiple people in a room,” Hanson says. “It builds a mental model of who you are, what you like and what you said. We’re getting to the point where it can hold an open-ended, open-domain conversation.” Hanson plans to roll out a $250 mass-market version in 2011 or 2012, with the same facial- and vocal-recognition capabilities. His goal is to provide a powerful testbed for researchers, while also harnessing AI algorithms to make a robot toy that’s actually fun for more than 15 minutes.

But for all of Nexi’s and Zeno’s social skills and painstaking simulation of emotional life, the bots are creatures of instinct, not introspection. Tracking software finds the human whom the robot is addressing, a keypad triggers a scripted response, and when you leave the room, they don’t imagine where you’ve gone, whether the conversation helped or hurt you, or how to over-throw your government. “It’s very difficult for an artificial intelligence to project to a physical sense,” says Kevin Warwick, a professor of cybernetics at the University of Reading in England. “A robot can think about eventualities, but it can’t think even one step ahead about the consequences of its decisions.”

There are, of course, researchers who foresee rapid progress in computational neuroscience leading to inevitable “strong AI,” or artificial intelligence that’s not simply finishing your sentence in a Google search box, but mimicking human thought. IBM’s Blue Brain Project, for one, is energizing doomsayers with its goal of creating a virtual brain, potentially as soon as 2019. Still, without a neurological map of our own sense of consequence or morality, the breakthroughs that would allow for a truly power-hungry or evil robot are nowhere in sight. Contemplating them is a little like debating the ethical pitfalls of unregulated teleportation. Until someone builds the Enterprise, why worry if Scotty is going to drunk-dial himself into your house?

Robots will not rise up en masse anytime soon. Nexi won’t be e-mailing Zeno the “exterminate all humans” file from R.U.R. to distribute among the world’s Roombas, Predators and assembly-line welding machines. It’s a fantasy, or, at best, a debate for another century. And like many robot fears, it threatens to drown out a more rational debate, one that stems from the fact that robots fall through nearly every legal and ethical crack. “If an autistic patient charges a robot and tries to damage it, how should the robot respond?” asks Lin, who is also planning to develop ethical guidelines for social healthcare bots. “Should it shut down? It’s an expensive piece of equipment—should it push back?” When the robots arrive in force, are we prepared for the collateral damage, both physical and psychological, they could inflict?

---

Zeno is more of a business plan than a stand-alone humanoid, an attempt by Hanson Robotics to channel the company’s breakthroughs in artificial skin and social-learning algorithms into a hybrid robot toy and dirt cheap research testbed. If Zeno catches on with kids, it could be the world’s biggest—and least controlled—experiment in human-robot interaction.

---

When our eyes see a robot, one that we think is autonomous—moving, acting, functioning under its own power—our mirror neurons fire. These same neurons activate when we watch another animal move, and neuroscientists suspect they’re associated with learning, by way of imitation. Mirror neurons could care less about a wax statue, or a remote-control drone. It’s the autonomous robot that lights the fuse, tricking the mind into creating a mechanical device as a living thing.

And yet, like many aspects of human–robot interaction, the full repercussions are unknown. Science-fiction writers may have spent a half-century theorizing about the long-term effects of living with robots, but science is only getting started. While the field of HRI goes about the business of collecting data and sorting out its methodologies, drawing solid conclusions can be impossible, or at least irresponsible. Take those mirror neurons, for example. Neuroscientists can watch them flip on, but the exact purpose of those neurons is still up for debate.

Another, more common example of the brain’s mysterious response to robots is often referred to as the uncanny valley—a poetic way of saying, “robots are creepy.” Proposed in a 1970 paper by roboticist Masahiro Mori, the uncanny valley describes a graph showing that humans feel more familiar with, and possibly more comfortable toward, humanoid machines. Until that, is the machine becomes too human-like, tripping the same psychological alarms associated with seeing a dead or unhealthy human. At that point the graph collapses, and then rises again with the response to a real human being, or, theoretically, a perfect android. Whether this is a distortion of our fight-or-flight instincts or something more complex, Mori’s word choice was important—the uncanny is not naked fear, but a mix of familiarity and fear, attraction and repulsion. It’s a moment of cognitive dissonance that the brain can’t reconcile, like encountering a talking Christmas tree, or a laughing corpse.

By academic standards, it’s evocative, exciting stuff, describing what appears to be a widespread phenomenon. Nexi’s unnerving YouTube clips seem like textbook examples, and the robot has plenty of unsettling company. The Japanese social bot CB2 (Child-robot with Biomimetic Body), with its realistic eyes, child-like proportions and gray skin, evokes near-universal horror among
Fear of a Bot Planet

It's not the hardware that makes the evil robot one of Western culture's most powerful myths. It's the software, the artificial intelligence (AI) that turns machines into monsters. Here are the most iconic examples of malevolent AI with the fears each inspired.

1600s
Golem of Prague
Taught us to fear: unstable artificial intelligence
In folk tales, the Golem of Prague was sculpted from river mud and animated with magic, but its design is robotic to the core—big, impossibly strong and emotionless. Its AI is also familiar in its limitations: The Golem floods a house when no one tells it to stop fetching water. In later versions of the myth, it loses its mind.

1818
Frankenstein's Monster
Taught us to fear: artificial genius
A doomed, romantic sociopath, the monster in Frankenstein had a whip-smart mind that was his own undoing. He learns to speak and read in months and to resent his creator just as quickly. Critics call this the world's most influential evil-robot story. Frankenstein refuses to build a mate, fearing a superior, malevolent race that would destroy mankind.

1921
Radius
Taught us to fear: organized robotic insurrection
Like Frankenstein's monster, the robots in the play R.U.R. are flesh-and-blood murderers. The difference is scale: These robots are mass-produced from factory-grown organs, and they succeed in wiping out the human race. The robot leader, Radius, doesn't mince words, saying, "I wish to be the master of people."

1950
The Machines
Taught us to fear: a less deadly but more secret insurrection
In his short story "The Evitable Conflict," science-fiction writer Isaac Asimov granted the machines control of the world economy. They proved zealous. Hollywood eventually supplied the melodrama, turning AI's quiet financial coup into the mass-house-arrest of mankind in the movie I, Robot.

1968
HAL 9000
Taught us to fear: AI-controlled systems
The singsong condescension in HAL 9000's voice should have been a warning sign. But by the time 2001: A Space Odyssey leaps from sci-fi to horror, it's too late—the AI jettisons the human crew members it considers to be a liability to the spacecraft's mission. Like Asimov's machines, HAL isn't malicious, just a little too smart for our own good.

1984
T-800/Skynet
Taught us to fear: networked, self-organizing AI
Skynet never appears on camera in The Terminator, but the movie's eponymous enforcer bears its message: The planet's not big enough for biological and artificial intelligence. We also don't see the advanced defense computer becoming self-aware. Instead, the movie shows the smoldering aftermath of war, giving an old myth its most powerful update.

1987
ED-209
Taught us to fear: armed, autonomous robots
In RoboCop, ED-209 has the cognitive powers of a very smart police dog and the firepower of an attack chopper. And, like dogs trained for violence, ED-209 sometimes bites the wrong person: In one of the movie's most memorable scenes, the security bot botches its own sales demo by gunning down an unarmed civilian.

1999
The Machines
Taught us to fear: everything in The Terminator, and robot slavers
The machines of The Matrix are a deliciously twisted race of AI. They turn prisoners into battery packs, craft vast virtual worlds to keep us occupied and, as evidenced by Agent Smith, are capable of abject hatred. The real horror of The Matrix (sequels aside) is the prospect of machines not only conquering mankind, but toying with our defeated species.

ILLUSTRATIONS BY SPLITINTOONE
bloggers and reporters. Another Japanese robot, KOBIAN, features a wildly expressive face, with prominent eyebrows and a set of fully formed, ruby-red lips. It, too, was instantly branded creepy by the Western press. The designers of those social bots were actually trying to avoid the uncanny—Asian labs are packed with photorealistic androids that leap headlong into the twitching, undead depths of Mori's valley.

But just as the Terminator scenario withers under scrutiny, the uncanny valley theory is nowhere near as tidy as it sounds. Based on those YouTube clips, I had expected my meeting with Nexi to be hair-currying. Instead, I can see my grin scattered across computer monitors in the Media Lab. Nexi's forehead-mounted, depth-sensing infrared camera shows my face as a black and gray blur, and the camera in its right eye portrays me in color. I watch as I slip from the monitors, Nexi's head and eyes smoothly tracking to the next face. I am not creeped out—I'm a little jealous. I want Nexi to look at me again.

“There are some very practical things that we do to make our robots not creepy,” Berlin says. The secret to Nexi's success, apparently, is in a mirror's reach of the robot: a slightly battered hardcover book titled The Illusion of Life: Disney Animation—required reading for the Personal Robots Group. “We’re making an animation, in real time,” Berlin says. Like many animated characters, Nexi's features and movements are those of exaggerated humanity. When it reaches for an object, its arm doesn’t shoot forward with eerie precision. It wastes time and resources, orienting its eyes, head and body, and lazily arcing its hand toward the target. Nexi is physically inefficient, but socially proficient.

How proficient? In interactions with hundreds of human subjects, including residents of three Boston-area senior centers, researchers claim that no one has run screaming from Nexi. Quite the opposite: Many seniors tried to shake the robot’s hand, or hug it. At least one of them planted a kiss on it. “It interacts with people in this very social way, so people treat it as a social entity in an interpersonal way, rather than a machine-like way,” Cynthia Breazeal, director of the Personal Robots Group, says. “In studies with Nexi, we've shown that if you have the robots behave and move in ways that are known to enhance trust and engagement, the reaction is the same as it is with people. You’re pushing the same buttons.”

That principle has proven true for CB2 and KOBIAN as well. The research leaders of both projects claim that the apprehension directed at their robots online and in the media never materializes in person. With the exception of one Thai princess, everyone who encountered CB2 liked it, according to Osaka University’s Minoru Asada. A Japanese newspaper brought a group of elderly to visit KOBIAN. They were “deeply pleased and moved,” Atsuo Takanishi, a professor of mechanical engineering at Waseda University, says, “as if the robot really had emotion.”

Even if the uncanny valley ends up being more of a shallow trench, one that’s easily leveled by actually meeting an android, the success of Nexi and company only raises a more profound question: Why do we fall so hard for robots? “It turns out that we’re vulnerable to attachment, emotionally, to objects. We are extremely cheap dates,” says Sherry Turkle, director of the MIT Initiative on Technology and Self. “Do we really want to exploit that?” Turkle has studied the powerful bond that can form between humans and robots such as Paro, an almost painfully cute Japanese baby-seal-shaped therapy bot that squirms in your arms, coos when caressed and reacts by sucking on a cabled pacifier. She has also documented assumptions of intelligence and even emotion reported by children playing with robotic dolls. The effect that Paro, a therapy bot that’s little more than an animatronic stuffed animal, had on senior citizens only reinforced her concerns. “Tell me again why I need a robot baby sitter?” Turkle asks. “What are we saying to the child? What are we saying to the older person? That we’re too busy with e-mail to care for those in need?”

To researchers like Turkle, the widespread deployment of social robots is as risky as it is inevitable. With some analysts estimating a $15 billion market for personal robots by 2015, the demand for expressive machines is expected to be voracious. At the heart of Turkle’s argument—a call for caution, essentially—is the fear of outsourcing human interaction to autonomous machines. Even more alarming are the potential beneficiaries of robotic companionship, from children in underfunded schools to seniors suffering from Alzheimer’s. Enlisting an army of robots to monitor the young and the elderly could be a bargain compared to the cost of hiring thousands of teachers and live-in nurses. But how will the first generation to grow up with robotic authority figures and friends handle unpredictable human relationships? Without more data, a well-intended response to manpower shortage could take on the ethical and legal dimensions of distributing a new and untested antidepressant.

One possible solution is to scale back the autonomy and use social bots as puppets. Huggable, another robot from MIT’s Personal Robots Group, is a teddy bear whose movements can be controlled through a Web browser. The researchers plan to use it to comfort hospitalized children; family members or doctors would operate it remotely. When I see Huggable, it’s actually a teddy bear skeleton. The furry coat, which will eventually be replaced with one that includes pressure- and touch-sensitive sensors, sits in a heap next to the bot as it fidgets. An open laptop shows the operator’s view through Huggable’s camera and a menu of simple commands, such as raising and lowering its arms, or aiming its head at my face.

For now, Huggable has no identity of its own. It’s a high-tech ventrilquist’s dummy channeling the voice of its operator, not a full-fledged social creature. In a recent paper...
The dangers of "nurture" in the design of robots are not simply compatible in their interactions with people...