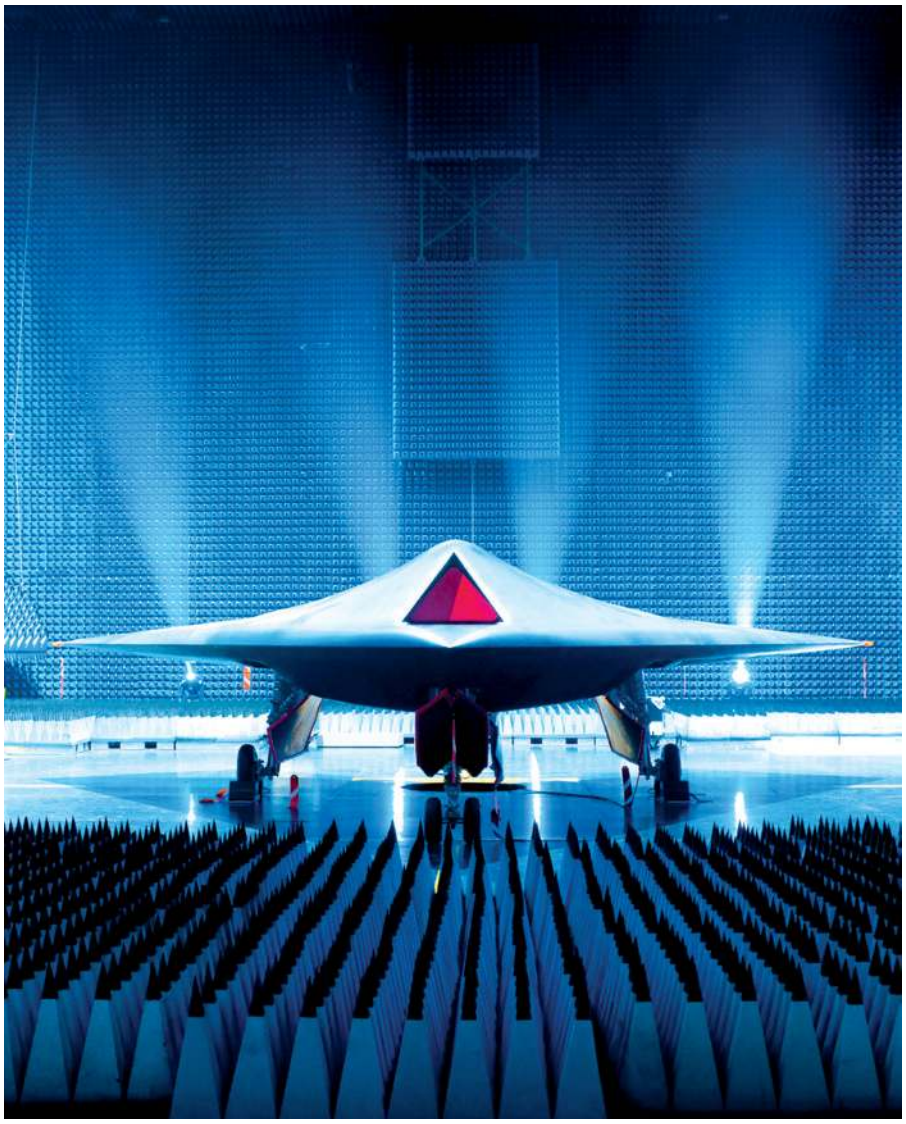
COMMENT



TONY GARNER/BAE

BAE Systems' Taranis drone has autonomous elements, but relies on humans for combat decisions.

Ethics of artificial intelligence

Four leading researchers share their concerns and solutions for reducing societal risks from intelligent machines.

STUART RUSSELL
Take a stand on AI weapons

Professor of computer science, University of California, Berkeley

The artificial intelligence (AI) and robotics communities face an important ethical decision: whether to support or oppose the development of lethal autonomous weapons systems (LAWS).

Technologies have reached a point at which the deployment of such systems is — practically if not legally — feasible within years, not decades. The stakes are high: LAWS have been described as the third revolution in warfare, after gunpowder and nuclear arms.

Autonomous weapons systems select and engage targets without human intervention; they become lethal when those targets include humans. LAWS might include, for example, armed quadcopters that can search for and eliminate enemy combatants in a city, but do not include cruise missiles or remotely piloted drones for which humans make all targeting decisions.

Existing AI and robotics components can provide physical platforms, perception, motor control, navigation, mapping, tactical decision-making and long-term planning. They just need to be combined. For example, the technology already demonstrated for self-driving cars, together with the human-like tactical control learned by DeepMind's DQN system, could support urban search-and-destroy missions.

Two US Defense Advanced Research Projects Agency (DARPA) programmes foreshadow planned uses of LAWS: Fast Lightweight Autonomy (FLA) and Collaborative Operations in Denied Environment (CODE). The FLA project will program tiny rotorcraft to manoeuvre unaided at high speed in urban areas and inside buildings. CODE aims to develop teams of autonomous aerial vehicles carrying out "all steps of a strike mission — find, fix, track, target, engage, assess" in situations in which enemy signal-jamming makes communication with a human commander impossible. Other ▶

countries may be pursuing clandestine programmes with similar goals.

International humanitarian law — which governs attacks on humans in times of war — has no specific provisions for such autonomy, but may still be applicable. The 1949 Geneva Convention on humane conduct in war requires any attack to satisfy three criteria: military necessity; discrimination between combatants and non-combatants; and proportionality between the value of the military objective and the potential for collateral damage. (Also relevant is the Martens Clause, added in 1977, which bans weapons that violate the "principles of humanity and the dictates of public conscience".) These are subjective judgments that are difficult or impossible for current AI systems to satisfy.

The United Nations has held a series of meetings on LAWS under the auspices of the Convention on Certain Conventional Weapons (CCW) in Geneva, Switzerland. Within a few years, the process could result in an international treaty limiting or banning autonomous weapons, as happened with blinding laser weapons in 1995; or it could leave in place the status quo, leading inevitably to an arms race.

As an AI specialist, I was asked to provide expert testimony for the third major meeting under the CCW, held in April, and heard the statements made by nations and non-governmental organizations. Several countries pressed for an immediate ban. Germany said that it "will not accept that the decision over life and death is taken solely by an autonomous system"; Japan stated that it "has no plan to develop robots with humans out of the loop, which may be capable of committing murder" (see go.nature.com/fwric1).

The United States, the United Kingdom and Israel — the three countries leading the development of LAWS technology — suggested that a treaty is unnecessary because they already have internal weapons review processes that ensure compliance with international law.

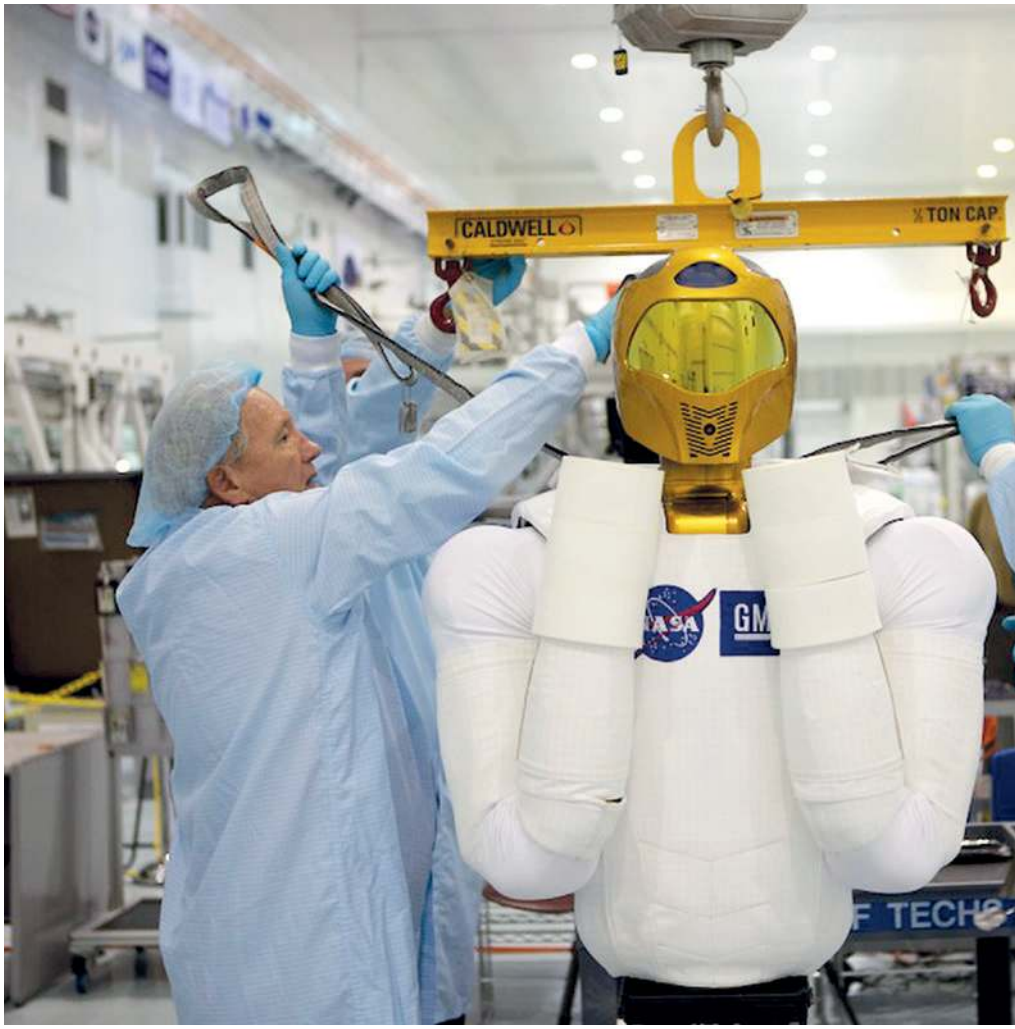
Almost all states who are party to the CCW agree with the need for 'meaningful human control' over the targeting and engagement decisions made by robotic weapons. Unfortunately, the meaning of 'meaningful' is still to be determined.

The debate has many facets. Some argue that the superior effectiveness and selectivity of autonomous weapons can minimize civilian casualties by targeting only combatants. Others insist that LAWS will lower the threshold for going to war by making it possible to attack an enemy while incurring no immediate risk; or that they will enable terrorists and non-state-aligned combatants to inflict catastrophic damage on civilian populations.

LAWS could violate fundamental principles of human dignity by allowing machines to choose whom to kill — for example, they might be tasked to eliminate anyone exhibiting 'threatening behaviour'. The potential for LAWS technologies to bleed over into peacetime policing functions is evident to human-rights organizations and drone manufacturers.

In my view, the overriding concern should be the probable endpoint of this technological trajectory. The capabilities of autonomous weapons will be limited more by the laws of physics — for example, by constraints on range, speed and payload — than by any deficiencies in the AI systems that control them. For instance, as flying robots become smaller, their manoeuvrability increases and their ability to be targeted decreases. They have a shorter range, yet they must be large enough to carry a lethal payload — perhaps a one-gram shaped charge to puncture the human cranium. Despite the limits imposed by physics, one can expect platforms deployed in the millions, the agility and lethality of which will leave humans utterly defenceless. This is not a desirable future.

The AI and robotics science communities, represented by their professional societies, are obliged to take a position, just as physicists have done on the use of nuclear weapons, chemists on the use of chemical agents and biologists on the use of disease agents in warfare. Debates should be organized at scientific meetings; arguments studied by ethics committees; position papers written for society publications; and votes taken by society members. Doing nothing is a vote in favour of continued development and deployment.



NASA's Robonaut 2 could be used in medicine and industry as well as space-station construction.

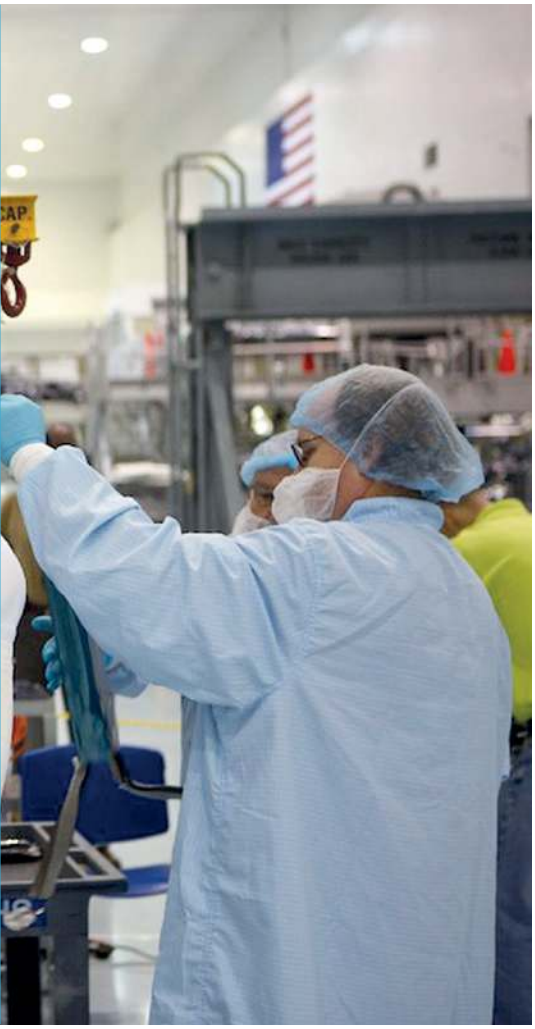
JOSEPH BIBBY/NASA

SABINE HAUERT

Shape the debate, don't shy from it

Lecturer in robotics, University of Bristol

Irked by hyped headlines that foster fear or overinflate expectations of robotics and artificial intelligence (AI), some researchers have stopped communicating with the



media or the public altogether.

But we must not disengage. The public includes taxpayers, policy-makers, investors and those who could benefit from the technology. They hear a mostly one-sided discussion that leaves them worried that robots will take their jobs, fearful that AI poses an existential threat, and wondering whether laws should be passed to keep hypothetical technology 'under control'. My colleagues and I spend dinner parties explaining that we are not evil but instead have been working for years to develop systems that could help the elderly, improve health care, make jobs safer and more efficient, and allow us to explore space or beneath the oceans.

Experts need to become the messengers. Through social media, researchers have a public platform that they should use to drive a balanced discussion. We can talk about the latest developments and limitations, provide the big picture and demystify the technology. I have used social media to crowd-source designs for swarming nanobots to treat cancer. And I found my first PhD student through his nanomedicine blog.

The AI and robotics communities need thought leaders who can engage with prominent commentators, such as physicist Stephen Hawking and entrepreneur–inventor Elon Musk, and set the agenda at international meetings such as the World Economic Forum in Davos, Switzerland. Public engagement also drives funding. Crowdfunding for JIBO, a personal robot for the home developed by Cynthia Breazeal, at the Massachusetts Institute of Technology (MIT) in Cambridge, raised more than US$2.2 million.

There are hurdles. First, many researchers have never tweeted, blogged or made a YouTube video. Second, outreach is 'yet another thing to do', and time is limited. Third, it can take years to build a social-media following that makes the effort worthwhile. And fourth, engagement work is rarely valued in research assessments, or regarded seriously by tenure committees.

> ***"Through social media, researchers have a public platform that they should use to drive a balanced discussion."***

Training, support and incentives are needed. All three are provided by Robohub.org, of which I am co-founder and president. Launched in 2012, Robohub is dedicated to connecting the robotics community to the public. We provide crash courses in science communication at major AI and robotics conferences on how to use social media efficiently and effectively. We invite professional science communicators and journalists to help researchers to prepare an article about their work. The communicators explain how to shape messages to make them clear and concise and avoid pitfalls, but we make sure the researcher drives the story and controls the end result. We also bring video cameras and ask researchers who are presenting at conferences to pitch their work to the public in five minutes. The results are uploaded to YouTube. We have built a portal for disseminating blogs and tweets, amplifying their reach to tens of thousands of followers.

I can list all the benefits of science communication, but the incentive must come from funding agencies and institutes. Citations cannot be the only measure of success for grants and academic progression; we must also value shares, views, comments and likes. MIT robotics researcher Rodney Brooks's classic 1986 paper on the 'subsumption architecture', a bio-inspired way to program robots to react to their environment, gathered nearly 10,000 citations in three decades (R. Brooks *IEEE J. Robot. Automat.* **2,** 14–23; 1986). A video of Sawyer, a robot developed by Brooks's company Rethink Robotics, received more than 60,000 views in one month (see go.nature.com/jqwfmz). Which has had more impact on today's public discourse?

Governments, research institutes, business-development agencies, and research and industry associations do welcome and fund outreach and science-communication efforts. But each project develops its own strategy, resulting in pockets of communication that have little reach.

In my view, AI and robotics stakeholders worldwide should pool a small portion of their budgets (say 0.1%) to bring together these disjointed communications and enable the field to speak more loudly. Special-interest groups, such as the Small Unmanned Aerial Vehicles Coalition that is promoting a US market for commercial drones, are pushing the interests of major corporations to regulators. There are few concerted efforts to promote robotics and AI research in the public sphere. This balance is badly needed.

A common communications strategy will empower a new generation of roboticists that is deeply connected to the public and able to hold its own in discussions. This is essential if we are to counter media hype and prevent misconceptions from driving perception, policy and funding decisions.

NATURE.COM
For more, see the *Nature* Insight on machine intelligence:
go.nature.com/eizewe

RUSS ALTMAN
Distribute AI benefits fairly

Professor of bioengineering, genetics, medicine and computer science, Stanford University

Artificial intelligence (AI) has astounding potential to accelerate scientific discovery in biology and medicine, and to transform health care. AI systems promise to help make sense of several new types of data: measurements from the 'omics' such as genomics, proteomics and metabolomics; electronic health records; and digital-sensor monitoring of health signs.

Clustering analyses can define new syndromes — separating diseases that were thought to be the same and unifying others that have the same underlying defects. Pattern-recognition technologies may match disease states to optimal treatments. For example, my colleagues and I are identifying groups of patients who are likely to respond to drugs that regulate the immune system on the basis of clinical and transcriptomic features.

CORBIS

Kirobo, Japan's first robot astronaut, was deployed to the International Space Station in 2013.

In consultations, physicians might be able to display data from a 'virtual cohort' of patients who are similar to the one sitting next to them and use it to weigh up diagnoses, treatment options and the statistics of outcomes. They could make medical decisions interactively with such a system or use simulations to predict outcomes on the basis of the patient's data and that of the virtual cohort.

I have two concerns. First, AI technologies could exacerbate existing health-care disparities and create new ones unless they are implemented in a way that allows all patients to benefit. In the United States, for example, people without jobs experience diverse levels of care. A two-tiered system in which only special groups or those who can pay — and not the poor — receive the benefits of advanced decision-making systems would be unjust and unfair. It is the joint responsibility of the government and those who develop the technology and support the research to ensure that AI technologies are distributed equally.

"AI technologies could exacerbate existing health-care disparities and create new ones."

Second, I worry about clinicians' ability to understand and explain the output of high-performance AI systems. Most health-care providers will not accept a complex treatment recommendation from a decision-support system without a clear description of how and why it was reached.

Unfortunately, the better the AI system, the harder it often is to explain. The features that contribute to probability-based assessments such as Bayesian analyses are straightforward to present; deep-learning networks, less so.

AI researchers who create the infrastructure and technical capabilities for these systems need to engage doctors, nurses, patients and others to understand how they will be used, and used fairly.

MANUELA VELOSO

Embrace a robot–human world

Professor of computer science, Carnegie Mellon University

Humans seamlessly integrate perception, cognition and action. We use our sensors to assess the state of the world, our brains to think and choose actions to achieve objectives, and our bodies to execute those actions. My research team is trying to build robots that are capable of doing the same — with artificial sensors (cameras, microphones and scanners), algorithms and actuators, which control the mechanisms.

But autonomous robots and humans differ greatly in their abilities. Robots may always have perceptual, cognitive and actuation limitations. They might not be able to fully perceive a scene, recognize or manipulate any object, understand all spoken or written language, or navigate in any terrain. I think that robots will complement humans, not supplant them. But robots need to know when to ask for help and how to express their inner workings.

To learn more about how robots and humans work together, for the past three years we have shared our laboratory and buildings with four collaborative robots, or CoBots, which we developed. The robots look a bit like mechanical lecterns. They have omnidirectional wheels that enable them to steer smoothly around obstacles; camera and lidar systems to provide depth vision; computers for processing; screens for communication; and a basket to carry things in.

Early on, we realized how challenging real environments are for robots. The CoBots cannot recognize every object they encounter; lacking arms or hands they struggle to open doors, pick things up or manipulate them. Although they can use speech to communicate, they may not recognize or understand the meaning of words spoken in response.

We introduced the concept of 'symbiotic autonomy' to enable robots to ask for help from humans or from the Internet. Now, robots and humans in our building aid one another in overcoming the limitations of each other.

CoBots escort visitors through the building or carry objects between locations, gathering useful information along the way. For example, they can generate accurate maps of spaces, showing temperature, humidity, noise and light levels, or WiFi signal strength. We help the robots to open doors, press lift buttons, pick up objects and follow dialogue by giving clarifications.

There are still hurdles to overcome to enable robots and humans to co-exist safely and productively. My team is researching how people and robots can communicate more easily through language and gestures, and how robots and people can better match their representations of objects, tasks and goals.

We are also studying how robot appearance enhances interactions, in particular how indicator lights could reveal more of a robot's inner state to humans. For instance, if the robot is busy, its lights may be yellow, but when it is available they are green.

Although we have a way to go, I believe that the future will be a positive one if humans and robots can help and complement each other. ■ SEE INSIGHT P.435