STUART RUSSELL

Banning Lethal Autonomous Weapons: An Education

Lethal autonomous weapons systems pose new and dangerous threats, but efforts to advocate for a ban demonstrate the complexities of finding international consensus.

Lethal autonomous weapons systems—commonly but misleadingly known as "killer robots"—are weapons systems that, once activated, can attack objects and people without further human intervention. With more than a dozen nations working to develop highly capable versions of them for use in the air, at sea, and on land, these weapons are not science fiction: they exist now, and they are already being used in some current conflicts.

Since 2014, the United Nations has held discussions around a treaty to ban autonomous weapons systems (AWS). So far, in addition to the UN secretary-general and the International Committee of the Red Cross, 30 countries have declared support for such a treaty. But the United States and Russia have combined forces to prevent any discussion of a legally binding instrument. Instead, in 2021 the United States called for a "non-binding code of conduct."

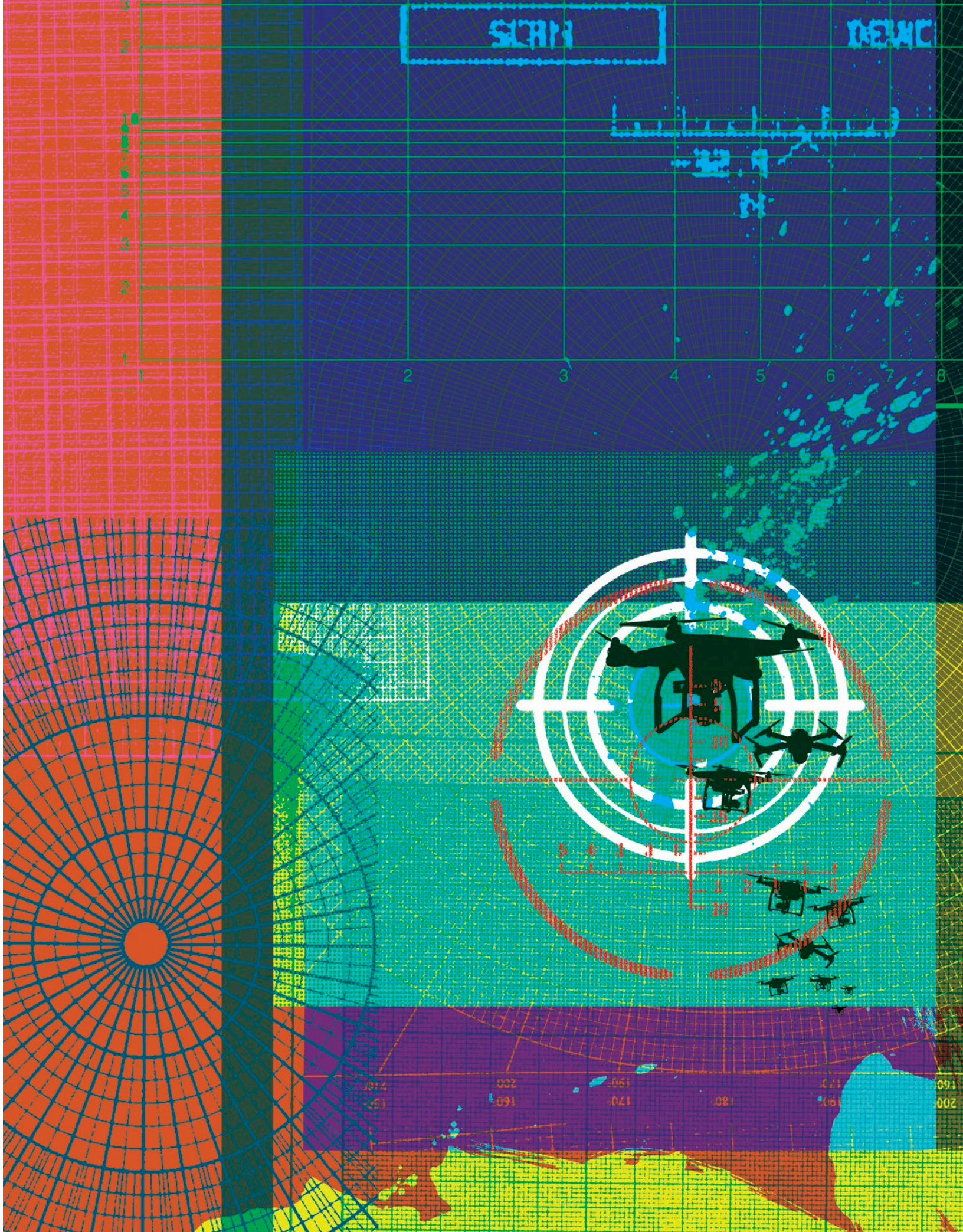
My involvement in the AWS policy discussion began in February 2013 when a puzzling email arrived from Human Rights Watch (HRW). I have studied artificial intelligence (AI) topics for 45 years and spent more than a decade working on verification for the Comprehensive Nuclear-Test-Ban Treaty. And I have been a member of HRW's Northern California committee for some time. For more than four decades, the organization had investigated atrocities around the world—atrocities committed by humans.

Now, HRW was asking me to support a new campaign to ban killer robots. The letter raised the possibility of children playing with toy guns being accidentally targeted by killer robots. It stated that robots would not be restrained by "human compassion, which can provide an important check on the killing of civilians."

I recovered from my initial confusion and replied to the email that perhaps we could start with a professional code of conduct for computer scientists—something like, *Do not design algorithms that can decide to kill humans.*

This admonition seemed to me an obviously sensible rule that any normal person would agree with. Yet I soon learned that "sensible" and "normal" are not words commonly associated with the geopolitical and diplomatic realm where arms control issues are discussed. In this arena, interests are competing and overlapping. Arguments based on merit play at best a secondary role, and can even hurt one's cause. Over the last eight years, I have slowly learned to navigate this world. In the beginning, I followed the standard policy script: giving dozens of presentations, including several at the United Nations, participating in hundreds of media interviews and events, and leading a delegation of scientists to the White House. But, as it became clear this was not enough, I soon tried other approaches, including initiating a petition signed by 30,000 people and originating a short film seen by millions.

In this article I will explain my education in this new arms control arena, but let me begin with some caveats. First, I'm not talking about banning *all* uses of AI in military applications. Some uses, such as better detection of surprise attacks, could be beneficial. Second, this is not about the general morality of defense research. For what it's worth, I think scientists have some obligation to help



those who are willing to die to protect them. Finally, I'm not talking about drones in the sense of aircraft that are remotely piloted by humans; they are not autonomous. Nor am I talking about technologies such as antimissile defense systems; they are autonomous but not lethal, since their targets are missiles and not humans.

After responding to HRW's email, I spent time learning about the governance of autonomous weapons. All weapons are governed in part by international humanitarian law, which includes the Geneva Conventions—in particular, the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects (better known, for obvious reasons, as CCW).

One of the main rules of international humanitarian law is the principle of distinction: one cannot attack civilians, and, by extension, one cannot use weapons that are by nature indiscriminate. A 2013 UN report warned that autonomous weapons might be indiscriminate—accidentally targeting civilians. From this warning came HRW's report, with its example of children being targeted because they're playing with toy guns.

Over time, I have come to think that this focus on accidental targeting was a strategic mistake, but it was one of the primary concerns that led to CCW's first discussion of autonomous weapons in 2014. Every year since, the countries involved in CCW have met to discuss the evolving capabilities of AWS and whether and how they should be controlled.

In 2015, I was invited to address the CCW meeting in Geneva as an AI expert. I had three jobs to do: clear up confusion about the meaning of "autonomy," assess the technological feasibility of autonomous weapons, and evaluate the pros and cons of using them. In my naïveté, this seemed like a chance to steer the discussion in a sensible direction.

Explaining autonomy didn't seem that difficult. To an AI researcher, the autonomy that matters for weapons is exactly the same kind we give to chess programs. Although we write the chess program, we do not decide what moves to make. We press the "start" button and the chess program makes the decisions. Very quickly it will get into board positions no one has ever seen before, and it will decide, based on what it sees and its own complex and opaque calculations, where to move the pieces and which enemy pieces to take.

That's precisely the UN definition: weapons that, once activated, can select and engage targets without further human intervention. There's no mystery, no evil intent, no self-awareness; just complex calculations that depend on what the machine's camera sees—that is, on information that is not available to the human operator at the time the weapon is sent on its mission.

The second question was feasibility—could these weapons be built with then-current technologies, and, if not, how long until it would be possible? For reasons I could not fathom, the arms control community—including HRW and a group of 20 Nobel Peace Prize winners—insisted that these weapons "could be developed within 20 to 30 years." In contrast, my robotics colleagues said "18 months, tops," and Britain's Ministry of Defence said some degree of autonomy is "probably achievable now" for some scenarios. Indeed, by 2015 all the component technologies for autonomous weapons already existed and would not be difficult to put together.

At the very least, an autonomous weapon requires a mobile platform. Even in 2015, there were already many options: quadcopters ranging from 3 centimeters to 1 meter in size; fixed-wing aircraft ranging from hobby-sized package delivery planes to full-sized missile-carrying drones; self-driving cars, trucks, and tanks; swarms of armed, unmanned boats; and even—if you must—skeletal humanoid robots. There were demonstrations of quadcopters catching balls in midair and flying sideways through vertical slots at high speed—even large formations of them filing through narrow windows and re-forming inside buildings. Nowadays, perfectly coordinated aerobatic displays of over 3,000 quadcopters are routine at corporate events.

Next, the machine must be able to perceive its environment. In 2015, the algorithms already deployed on self-driving cars could track moving objects in video, including human beings and vehicles. Autonomous robots could already explore and build a detailed map of a city neighborhood or the inside of a building.

The machine must also have the ability to make tactical decisions. These might resemble the ones demonstrated by AI systems in multiplayer video games or in self-driving cars. In many senses, however, designing a weapon is easier for the simple reason that a self-driving car cannot make any mistakes, but an AWS that works 80% of the time is perfectly adequate according to military standards. (And if failure were a concern, it could be overcome by sending three AWS instead of just one.)

The final consideration is lethality. Some weapons were already available on remotely piloted drones, including vision-guided missiles, gyro-stabilized machine guns, and the 51-pound explosive carried by Israel's Harpy 2 loitering missile. Many other lethal technologies could easily be adapted to work on an autonomous platform.

After discussing feasibility, I turned to the pros and cons: Should countries develop and deploy autonomous weapons or should nations ban them? One commonly cited benefit of autonomy is that wars fought between robot armies might drastically reduce the risks to human combatants. But if that were true, we could also settle wars

by playing tiddlywinks. In the real world, I think, wars end when the level of death and destruction becomes untenable for one or both sides.

Perhaps the most serious argument in favor of autonomous weapons is the claim that they protect civilians. At the CCW meetings, the American and British delegations generally contend that autonomous weapons can "reduce risks to civilians in military operations, by … automating target identification … to improve speed, precision, and accuracy." This is an extension of the argument for remotely piloted drones; it requires AI systems that are better than humans at recognizing legitimate targets, which was probably not possible in 2015, although capabilities have advanced since. The problem with this argument is that it also implies that autonomous weapons will be used very much like drones. If they are not—if autonomous weapons are used more often, by different parties, against different targets, with different goals, or in less clear-cut settings—then civilian casualties could be far greater.

For this reason, I think that the emphasis on the issue of targeting accuracy and discrimination has been misguided—on both sides of the debate. The concern diverts attention from the big picture, which is that autonomous weapons will completely change the nature of warfare, the balance of power among nations and nonstate actors, and the viability of the right to "security of person" enshrined in the Universal Declaration of Human Rights.

The list of commonly raised objections to autonomous weapons is a long one. Some are quite practical, such as the fact that autonomous weapons might be subject to cyber-infiltration, causing them to turn against their owners once a war started, or they could accidentally escalate a conflict if a false alarm led to real, automated retaliation. Both cyber-infiltration and escalation are already taken seriously by military planners.

Campaigners have also raised legal arguments, such as the "accountability gap" that arises when AI systems commit atrocities. But proponents of autonomous weapons contend that there is no new gap here between criminal intent and criminal negligence on the part of the humans who launch the attack.

Finally, there are arguments about morality and honor. International humanitarian law includes an explicitly moral element called the Martens Clause, which says that "in cases not covered by the law in force, the human person remains under the protection of the principles of humanity and the dictates of the public conscience." One can see echoes of this moral principle in various public statements: for example, António Guterres, the UN secretary-general, stated in 2019, "Machines with the power and discretion to take lives without human involvement are politically unacceptable, morally repugnant, and should be prohibited by international law." Moral opposition to autonomous weapons has also come from unexpected quarters: in a surprise move in a 2016 debate at the World Economic Forum, Roger Carr, chairman of the defense contractor BAE Systems, stated that delegating kill decisions to machines was "fundamentally wrong" and pledged that his company would never allow it. And in 2017, Paul Selva, vice chairman of the Joint Chiefs of Staff in the United States, told Congress, "I don't think it's reasonable for us to put robots in charge of whether or not we take a human life."

There is no question that there is a sense of honor among soldiers. In meetings with high-level military officers from several countries, I have been struck by how seriously they take their responsibility for life and death. And of course, they understand that they could one day be on the receiving end of attacks by autonomous weapons.

One commonly cited benefit of autonomy is that wars fought between robot armies might drastically reduce the risks to human combatants. But if that were true, we could also settle wars by playing tiddlywinks.

I didn't believe, however, that arguments based on morality and honor alone would sway the governments who are key to decisions about these weapons—especially when they distrust the morality and honor of all the other governments. So the final question I explored while preparing for the CCW meeting was the future evolution of autonomous weapons. What kinds of weapons would AI enable, how would they be used, and how would that change war itself?

In short, rather than appealing to higher principles, I hoped to appeal to national self-interest.

It seemed to me that AI, by removing the human element, would enable a lethal unit to be far smaller, cheaper, and more agile than a tank or an attack helicopter or a soldier carrying a gun. A lethal AI-powered quadcopter could be smaller than a tin of shoe polish, and if it carried just three grams of explosive, it could kill a person at close range.

It's not hard to imagine that eventually, a weapon like this could be mass-produced very cheaply. And, to continue this speculative scenario, a regular shipping

container could hold a million of them. Because, by definition, no human supervision is required for each weapon, they could all be sent to do their work at once. The endpoint, I believe, would be that autonomous weapons become cheap, selective, scalable weapons of mass destruction (WMDs). Clearly, it seemed to me, this would be a disaster for international security—and thus, the self-interest of my governmental and diplomatic audience.

However, after my presentation in Geneva, I found myself in the unusual position of being extremely popular with the ambassadors from Cuba, Pakistan, and Venezuela, but not with the Americans and British. Their disgruntlement came, I suspect, from what they saw as *realpolitik*: the utilitarian need to maintain military superiority over any potential enemies who would develop AI weapons.

Although American and British opposition to a treaty has stymied progress in Geneva, my hope remains that these countries will eventually understand that the need to avoid creating a new category of cheap, scalable weapons of mass destruction is an argument that they in fact have already accepted. In 1966, a coalition of American biologists and chemists wrote to President Lyndon Johnson explaining that biological weapons, once perfected, would become cheap, widespread weapons of mass extermination. Johnson's successor, Richard Nixon, became convinced that these weapons would ultimately reduce American security, leading him to unilaterally renounce biological weapons in 1969. The United Kingdom, for its part, helped initiate negotiations on an international treaty to ban them, which became the Biological Weapons Convention.

Although there are ongoing debates over the definition of weapons of mass destruction, it seems to me self-evident that if someone can type a command, press "return," and wipe out a million people, that's a weapon of mass destruction. The security implications of small antipersonnel AWS, however, may be even greater than those of other WMD categories. Properly programmed, a swarm could wipe out, say, all the males aged between 12 and 60 in a city, or all members of an ethnic or religious group. And unlike nuclear weapons, the swarm would leave no radioactive crater, nor would it ruin valuable real estate. But, as others have written, the bigger difference between AWS and nuclear weapons is that the former are scalable: conflicts can escalate smoothly from ten to a thousand to a hundred thousand casualties with no identifiable calamitous threshold being crossed.

After the 2015 Geneva meeting, it was clear to treaty proponents within the AI community that we would need more than PowerPoint presentations to win this argument. In July 2015 we launched an open letter calling for a ban; 4,667 AI researchers signed, including almost the entire leadership of the field; they were joined by nearly 27,000 other signatories. Media articles appeared around the world. Even the *Financial Times*—not exactly the peaceniks' house journal—supported the ban, calling autonomous weapons "a nightmare the world has no cause to invent."

I also came to realize that any attempt to debate the issue of lethal autonomous weapons in the public realm is complicated by widespread, pre-existing misconceptions. The media persist in associating autonomous weapons with rampaging *Terminator*-style robots, which misleads the public into thinking that autonomous weapons are science fiction. The *Terminator* theme also makes people think that the problem is Skynet—the global software system that controls the Terminator robots—becoming conscious, hating humans, and trying to kill us all.

Any attempt to debate the issue of lethal autonomous weapons in the public realm is complicated by widespread, pre-existing misconceptions.

Of course, a conscious, malicious Skynet has never been a real problem, but that narrative clearly needed a counterargument. I became convinced that we needed to visually explain the problem with autonomous weapons. In 2017, with the help of writers and filmmakers at the Space Digital film and digital effects company and funding from the Future of Life Institute, we made a film called *Slaughterbots*. It had two storylines: one, a sales pitch by the CEO of an arms manufacturer, demonstrating the tiny quadcopter and its use in selective mass attacks; the other, a series of unattributed atrocities targeting, among others, the US Congress and university students in several countries.

The film premiered at the CCW meeting in November 2017. The reactions elsewhere were mostly positive: the film soon had millions of views on the web, and an article on CNN called it "the most nightmarish, dystopian film of 2017." The film had another pleasing side effect: still images from *Slaughterbots* gradually began to replace Terminators in media illustrations of the autonomous weapons issue.

Many of my AI colleagues thought the CEO's presentation was real, not fictional, which says something about where the technology stood in 2017. At the CCW meeting, on the other hand, the Russian ambassador responded to my presentation by angrily asking me: "Why are we discussing science fiction? Such weapons cannot exist for another 25 or 30 years!"

Earlier that year, however, a government-affiliated manufacturer in Turkey had revealed its latest product: the Kargu drone. The Kargu-2, its current version, is a multicopter the size of a dinner plate that is capable of going 90 miles per hour and carrying 3 pounds of explosives—enough to destroy vehicles and damage buildings, as well as kill and injure people. According to a 2021 UN report, Kargus were used in Libya in 2020 when retreating members of one faction were autonomously "hunted down and remotely engaged." The problem here was not accidental targeting of civilians, but deliberate targeting of combatants in full retreat (violating the rule of military necessity)—all in the context of a complete UN embargo on arms sales.

Fully autonomous weapons are unfortunately a reality, and opponents and proponents of these systems now find themselves at an unstable impasse—unstable because the technology is accelerating. Many countries are in favor of a ban, as are the European Parliament, the United Nations, the Non-Aligned Movement, hundreds of civil society organizations, and, according to one recent poll, 61% of adults across 28 countries. And yet, efforts to ban AWS are at a standstill because the American and Russian governments, supported to some extent by Britain, Israel, and Australia, argue that a ban is unnecessary.

When countering such intractable opposition, continuing to state one's own position more energetically is unhelpful—especially when the others have all the cards. Instead, a small group, convened by Massachusetts Institute of Technology physicist Max Tegmark in late 2019, decided to explore the possibility that a reasoned and collegial discussion could lead to a better outcome. We met at Max's house in Boston. The participants—advocates as well as opponents of autonomous weapons—included AI researchers and experts from military, arms control, and diplomatic backgrounds.

After some discussion and debate, during which the group seemed to be getting nowhere, we began to consider compromise solutions, such as a limited ban that would require a minimum weight and explosive payload size in order to rule out small antipersonnel weapons.

There's an interesting precedent for this called the St. Petersburg Declaration of 1868. The declaration's origins seem almost quaint today: a Russian engineer had invented a musket ball that exploded inside the body, and Russian diplomats soon realized such an inhumane weapon could set off an arms race. They convened a meeting, and the resulting declaration banned exploding ordnance below 400 grams—a ban that holds, at least approximately, to this day.

A similar ban on small antipersonnel AWS could eliminate swarms as weapons of mass destruction. But it would allow the major powers to keep developing their autonomous submarines, tanks, and fighter aircraft, thereby preventing strategic surprise.

A St. Petersburg-type agreement would be better than nothing, especially given the failure of the latest round of CCW negotiations in 2021. It would be far from ideal, however, as the principle that machines can decide whom to kill would de facto become widely accepted. Whereas a total ban maintains the moral stigma and the universal principle, a ban on only smaller-sized weapons will come under constant pressure as manufacturers vie to develop cheaper and more agile weapons. Moreover, a black market is likely to emerge. I think the moral simplicity of the bans on chemical and biological weapons is part of why they've been quite effective—for example, they ban irritant and anesthetic chemicals in war, even though these are allowed in many countries for domestic policing.

What's next? Inevitably, the diplomatic action will move to the UN General Assembly, where unanimity is not required for progress. (For example, the Comprehensive Nuclear-Test-Ban Treaty was adopted by the General Assembly in 1996 after it failed to progress within the consensus-based Conference on Disarmament.) It may also proceed outside the UN umbrella, as happened with the Antipersonnel Landmines Convention, which got its start 30 years ago when six nongovernmental organizations got together to build an international movement to monitor landmine use and advocate for a ban on the weapons.

Today, an important component of my strategy is getting the AI and robotics communities, represented by their professional societies, to take a position. The 400,000-strong Institute of Electrical and Electronics Engineers is in the process of developing a policy position, and I am starting to put together an autonomous weapons task force for the Association for Computing Machinery, the main professional society for computer science. We are planning to organize debates at scientific meetings, convene ethics committees to study the arguments, and encourage position papers. Some members of our profession protest that they don't have control over how their inventions are used; but it is clear, I think, that doing nothing is a vote in favor of continued development and deployment.

National and international policy advocacy, public outreach, and small-group expert consensus have all yielded little progress so far. Meanwhile, the technology driving autonomous weapons systems continues to advance. It is time for the AI community to step up, just as physicists stood against nuclear weapons, chemists against chemical weapons, biologists against biological weapons, and doctors did around their involvement in executions. Historically, the voice of scientists has mattered on such issues. And so I encourage my colleagues to join me in the exacting, detailed, confusing, and often frustrating work of allowing human beings to live their lives in relative security.

Stuart Russell *is a distinguished professor of computer science at the University of California, Berkeley, and director of the Center for Human-Compatible Artificial Intelligence and the Kavli Center for Ethics, Science, and the Public.*