"Hands Up, Don't Shoot!" HRI and the Automation of Police Use of Force

Peter Asaro
School of Media Studies, The New School
Center for Information Technology Policy, Princeton University
Center for Internet and Society, Stanford Law School

This paper considers the ethical challenges facing the development of robotic systems that deploy violent and lethal force against humans. While the use of violent and lethal force is not usually acceptable for humans or robots, police officers are authorized by the state to use violent and lethal force in certain circumstances in order to keep the peace and protect individuals and the community from an immediate threat. With the increased interest in developing and deploying robots for law enforcement tasks, including robots armed with weapons, the question arises as to how to design human-robot interactions (HRIs) in which violent and lethal force might be among the actions taken by the robot, or whether to preclude such actions altogether. This is what I call the "deadly design problem" for HRI. While it might be possible to design a system to recognize various gestures, such as "Hands up, don't shoot!," there are many more challenging and subtle aspects to the problem of implementing existing legal guidelines for the use of force in law enforcement robots. After examining the key legal and technical challenges of designing interactions involving violence, this paper concludes with some reflections on the ethics of HRI design raised by automating the use of force in policing. In light of the serious challenges in automating violence, it calls upon HRI researchers to adopt a moratorium on designing any robotic systems that deploy violent and lethal force against humans, and to consider ethical codes and laws to prohibit such systems in the future.

Keywords: violence, use of force, deadly force, police, HRI, automation, robocop

1. Introduction: A Deadly Design Problem

Recently, there has been a growing number of robotic systems introduced for use in law enforcement, security, and policing.[1] Some of these robots feature weapons such as tasers and tear gas, which could be used against people.[2]

[1] Dubai police forces have already obtained policing robots designed to interact with the public (https://www.rt.com/news/253529-police-robot-dubai-robocop). A police patrol robot has been developed by a Silicon Valley company, Knightscope (http://knightscope.com/about.html), and a South Korean company has been testing prison guard robots since 2012 (http://www.digitaltrends.com/cool-tech/meet-south-koreas-new-robotic-prison-guards).

[2] The design firm Chaotic Moon demonstrated a taser-armed drone on one of its interns at SXSW in 2014 (http://time.com/19929/watch-this-drone-taser-a-guy-until-he-collapses), while in the state of North Dakota, a bill designed to required warrants for police to use drones, and which originally prohibited arming police drones, was later amended to permit "non-lethal" weaponization, including tasers and teargas before being passed in August, 2015 (https://www.washingtonpost.com/news/the-switch/wp/2015/08/27/police-drones-with-tasers-it-could-happen-in-north-dakota).
A South African company, Desert Wolf, is marketing their Skunk drone, armed with teargas pellet guns, to mining companies to deal with striking workers (http://www.bbc.com/news/technology-27902634). The police department in Lucknow, India has already obtained five drones designed to disperse pepper spray for controlling crowds (http://fusion.net/story/117338/terrifying-

Additionally, there is growing use of intelligent sensors by law enforcement agencies that go beyond simple surveillance and toward automatic enforcement by collecting and processing data automatically, such as speeding and red-light traffic cameras, automatic license-plate readers,[3] and automated facial-recognition and biometrics.[4] While most of the recently developed police robots are remotely operated, rather than autonomous, and most are not weaponized,[5] research continues into increasingly autonomous patrol robots with a clear potential for being weaponized.

A recent report from Amnesty International raises concerns over how the use of robotic weapons, or weaponized robots, in policing could threaten human rights.[6] It follows a report from the United Nations Special Rapporteur on Extrajudicial, Summary and Arbitrary Executions that raised similar concerns.[7] The development of autonomous weaponized police robots that could potentially violate human rights raises a set of ethical and legal questions for human-robot interaction (HRI) design as a profession. I call this the "deadly design problem" for HRI: How, or whether, to design robotic systems that could deploy violent and lethal force against humans?

In this paper, I will examine the ethical and legal problems facing HRI designers who might be asked to build a law enforcement robot that could deploy violent force or lethal force against a person. What is their responsibility to uphold international human rights, given the kinds of technical problems such a deadly HRI design problem presents? Beyond the HRI community, I believe it is useful to take this design perspective in order to understand both the nature of how an engineer might view the problem, and to make clear how technical issues relate to the moral and legal requirements for the use of violent and lethal force by humans. Such an investigation will also, I hope, advance the discussion of these issues within the HRI community. In particular, should HRI designers even consider the use of violent or lethal force as being among the potential actions taken by a robot in its interactions with humans? I conclude that HRI professionals should adopt a moratorium on the design of systems capable of using force autonomously, should ensure that human operators have meaningful control over any robotic system capable of deploying violent and lethal force against humans, and should consider permanently banning such systems in the future.

I will proceed by examining four key design problems for developing robots capable of using violent or lethal force, which also raise moral and legal questions. The first problem is deciding which legal standards to implement or adhere to in the design. Given the extreme variations in laws and policies from state to state, and police department to police department within the United States and internationally, the choice of standards is a critical design question. Given the fact that none of the current standards in use in the United States meet the minimal international standards set by the United Nations Declaration on Human Rights, or the Guidelines issued by the UN Human Rights Council for the use of force and firearms, the choice of standards is also a moral and legal issue.

---

pepper-spray-drones-will-be-used-to-break-up-protests-in-india). Documents obtain from a FOIA by EFF.org in 2013 revealed that the US Customs and Border Patrol contemplated whether non-lethal weapons could be mounted on their unarmed predator drones for "immobilizing" suspicious persons (http://www.slate.com/blogs/future_tense/2013/07/03/documents_show_customs_and_border_protection_considered_weaponized_domestic.html).

[3] See, e.g., https://www.eff.org/deeplinks/2015/10/license-plate-readers-exposed-how-public-safety-agencies-responded-massive

[4] Kelly Gates (2011). *Our biometric future*. New York: NYU Press.

[5] After this paper was submitted for review, police in Dallas, Texas used a remote-operated bomb disposal robot, armed with a remotely triggered bomb, to kill a man suspected of killing multiple police officers during a #BlackLivesMatter protest (See McCarthy, Simone (2016). "What does Dallas's 'bomb robot' mean for the future of policing?," *Christian Science Monitor*, July 9, 2016, http://www.csmonitor.com/USA/USA-Update/2016/0709/What-does-Dallas-s-bomb-robot-mean-for-the-future-of-policing). In 2014, police in Albuquerque, New Mexico used a remote operated robot to deploy tear gas against a suspect who was barricaded in a building (Graham, David A. (2016). "The Dallas Shooting and the Advent of Killer Police Robots," *The Atlantic*, July 8, 2016, https://www.theatlantic.com/news/archive/2016/07/dallas-police-robot/490478).

[6] Amnesty International, Autonomous weapons systems: Five key human rights issues for consideration," April 10, 2015, https://www.amnesty.org/en/documents/act30/1401/2015/en

[7] Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions, Christof Heyns, to the Human Rights Council, A/HRC/23/47, April 9, 2013, para. 72, http://www.ohchr.org/Documents/HRBodies/HRCouncil/RegularSession/Session23/A-HRC-23-47_en.pdf

The second, and perhaps most difficult, design problem is to determine when it is appropriate to use force. This involves both the technical challenge of actually identifying those situations in which the use of force could be deemed legally acceptable and *necessary*, predicting when the use of force would be effective, as well as determining the individuals against whom violence might be appropriately directed—also known as *discrimination* in the use of force. Closely related is the third problem is to decide how much force is appropriate in the given circumstances, and when and how to escalate the use of force—also known as *proportionality* in the use of force. Again, there are questions of which legal standards to conform to, but also much more challenging technical issues involving how to meet those requirements, given that they may demand explicitly *human* judgments. Finally, the fourth problem involves questions of accountability—who should be responsible for the actions taken by this robot, how are they to be held accountable, and how can the design of the system make this clear?

I have several aims in examining these four design problems. In part, I wish to show that each of these design problems is simultaneously and necessarily technical, moral, and legal problems. In part, I also wish to show that in each case the technical solutions pose extreme engineering challenges and carry unavoidable risks and may not be solvable in the foreseeable future, or ever. Given either or both of these conclusions, I will argue that HRI designers should, at the very least, exercise extreme precaution in developing or deploying systems capable of violence, should be deeply skeptical of any purported technological solutions to any of these problems, and should think carefully and critically about any purported standards for evaluating such systems. And finally, I will conclude that there are strong technical, moral, and legal reasons for HRI researchers, and roboticists, more generally, to adopt a principle, as part of their professional code of ethics, to refrain from developing autonomous robots that are designed to deploy violent or lethal force against people, to ensure meaningful human control over any robots capable of deploying such force, and to consider supporting a permanent legal prohibition on such systems.

Before I begin, I think it is important to define some basic terms and concepts. As will become clear in the analysis below, "violence" means more than simply causing harm to a person, and will be defined as a meaningful act involving the use of physical force aimed at expressing an intention to assert control, as well as an exercise of that control, of one agent or group of agents over another. By "meaningful act" I mean both that it is intentional (rather than an accident) and that it means to cause a degree of harm to an individual through physical force. This definition comports with the general understanding of "use of force" recognized by law enforcement, and criminal and civil law. That is, "violence" includes both a "force" component and a symbolic or intentional component, meaning that the harm done has a purpose and was meant to be done. Of course, the actual harm caused by a violent act may be different than what was intended. The concept of exercising "restraint" over another person is a different one, and while it may involve violence, it may also be exercised in a non-violent way, or by means of threats including threats of violence.

There are subtly different ways to more narrowly define the use of force, and I have carefully selected the phrase "use of violent and lethal force." In particular, "use" implies that a robot is deploying force in relation to a specific goal, rather than accidentally or incidentally in relation to another goal. Thus, when a self-driving car runs over a person on its way to a destination, it is not "using force" against them unless it had a specific goal to cause injury through hitting the person as an explicit part of its action plan. Such harms are accidents, rather than violence. Violence, in this sense, is an intentional harm. Since the beginning of industrial robotics, robots have been dangerous and capable of causing real injuries and harms. Indeed, much HRI and robotics research has been devoted to obstacle avoidance, compliant actuators, and other technologies that can reduce these dangers. But such physical harms are not violent unless they also fulfill an intention or plan to do such harm.

For the most part, I will be concerned with bodily injuries as physical harms, as opposed to non-physical forms of harm, such as emotional, psychological, reputational, financial, and so forth. I leave it to future work to determine if the same or similar arguments extend to such harms. Causally speaking, even acts that direct only a small amount of force can cause great harms (e.g., triggering a gun or some causal system set to release a much larger violent force, or pushing someone off a cliff), as well as threaten further harms (e.g., violently poking a finger), and could thus be considered violent force even when the physical forces involved are quite small. The distinction between violent force and lethal force is somewhat more complex, because it relies on both intents and effects (or

expectations of effects).[8] This will be discussed further in the discussion below of the proportionate use of force. Suffice it to say here that the use of both violent force and lethal force by robots poses a range of shared design problems, even if there are some important differences.

The religious and philosophical traditions of the world have developed a number of different systems of morality and ethics, though they largely agree that violence and killing are generally prohibited. Laws and policies, as enacted and adopted by local, state, federal and international bodies, represent both political compromises and, in some sense, a synthesis of the main moral and ethical systems. That is, laws and policies both aim to be moral but also inevitably represent combined elements of various moral perspectives and theories. As such, it is important to recognize this when implementing these standards and to avoid compromising one moral perspective at the expense of others, even when it may be advantageous from a technical perspective to do so. That said, I will consider human rights as implying both legal and moral duties on HRI designers.

2. Which Standards for the Use of Violent Force and Lethal Force Should Apply to Robots?

Roboticists and HRI designers generally aim to reduce the risks of potential harms caused by their systems. They thus face a deadly design problem once they start to consider designing a system capable of using violent and lethal force against humans, and thus *deliberately causing harm to the people it interacts with*. According to social norms, moral systems, and laws, it is understood that the use of force is only acceptable in certain exceptional circumstances (i.e., in self-defense, or in the defense of another person). But the various social, moral, and legal standards do not always agree on which circumstances those are, what reasons justify the use of violent and lethal force, and what conditions apply to the initiation and escalation of violent and lethal force.

In technological terms, it is already possible to design a robotic system that is capable of targeting and firing a weapon, such as a gun or taser, with some degree of accuracy. Far more challenging is designing a system that only uses force when it is legally *necessary*, one that uses that force *discriminately*, and one that uses force *proportionately*. Beyond the technical challenges of building a system that can adhere to given rules for the necessity of the use of force, there are also serious questions about which standards or set of rules ought to be adhered to, or "built into" the system, and how those rules ought to be interpreted in actual situations.

2.1. United States Standards for the Use of Force

If asked to build a law enforcement robot for use by police in the United States, what use of force standards should a responsible HRI designer use as a design constraint for their robot to adhere to? Can a robot use force against a fleeing suspect? What kind of warning is necessary prior to the use of force? What kind of threat must a suspect pose in order to warrant the use of violent or lethal force? What other actions must the robot attempt before resorting to violent or lethal force?

As a recent Amnesty International report[9] makes clear, there is great variety in local and state policies and laws governing the use of violent and lethal force by police. At the federal level, Supreme Court decisions have set constitutional law standards for the use of violent and lethal force, while the Department of Justice has issued its own guidelines, but there is no specific federal legislation in place. Most state and local policies actually fail to meet either or both of the federal standards established by the Supreme Court and Department of Justice. As a designer,

[8] There is an ongoing debate regarding the distinction between lethal and non-lethal weapons. This is not a simple matter of probability, as so-called non-lethal weapons, such as tasers, frequently kill people, while in practice, guns often miss their targets, and the percentage of shots fired that are lethal can be relatively low (see: Wright, Steve (2001). "The role of sub-lethal weapons in human rights abuse," *Medicine, Conflict and Survival*, 17(3), pp. 221-233). Guns might be classified as lethal weapons, with the potential to kill, but also can be aimed low to try to avoid having lethal effects. Some uses of force, such as choking, can be calibrated, but may still result in unintentional death. For these reasons, it is better to use the term 'less-lethal weapons' instead of 'non-lethal.' Lethal weapons are those that can be reasonably expected to have a significant probability of causing death.

[9] Amnesty International (2015). Deadly Force: Police Use of Lethal Force in the United States, Report, June 18, 2015. https://www.amnestyusa.org/sites/default/files/aiusa_deadlyforcereportjune2015.pdf

should one design different systems for each state and local jurisdiction? Or choose one, or both, of the federal standards?

More distressing, however, is that the established laws or policies in the United States at all levels and jurisdictions *fail to conform to international standards* for the use of violent and lethal force by police. This includes failures to meet the minimal standards established by the United Nations Human Rights Council. These failures are as complete and far-reaching as they are distressing. That is to say that some states fail to establish any laws or policies regarding police use of violent and lethal force, while many others establish far lower standards than what is called for by international law, and even federal standards fail to meet the minimal international standards.

These shortcomings range from permitting the use of force to gain compliance with "lawful orders;" to using lethal force against fleeing individuals even when they pose no significant risk to cause harm; to permitting lethal force as a first resort rather than last; to failing to establish policies and procedures for documenting the use of force and discharge of firearms; to failing to establish inquiries into police actions resulting in death and serious injury; to failing to provide oversight mechanisms for monitoring police use of force and training. All of these are failures to meet the international guidelines, which only permit the use of force when there is an immediate threat of grave bodily harm or death and can only be averted by applying violent or lethal force against the individual posing the threat. This means that it is unacceptable to use force simply to achieve compliance with orders, prevent a suspect or prisoner from fleeing (unless they pose a grave and imminent threat), and there are further requirements to use the least amount of force necessary to prevent the imminent harm, as well as a requirement to give warning before force is used, when possible. Furthermore, there are federal reporting requirements for police departments to report incidents of the use of lethal force that have largely gone unenforced.

2.2. Challenges for Design Standards

The first conclusion to draw is that it might be necessary to design different systems for the use of force in different jurisdictions. One option might be to allow the police departments that acquire the robots to decide which use of force standards to utilize, though that would mean that designers would have to develop all of the possible variations. The second and more challenging issue is that the designers must decide whether to develop robots that may meet local standards but fail to meet international standards. Building to local standards might be permissible where those standards are more restrictive than the international standards but would be legally and morally problematic where they are less restrictive.

This raises a question of the ethical responsibility of designers—should they design systems that they know violate one set of standards, even if they meet other, weaker standards? Do they have an obligation to inform the users of the system of such situations? Do the users bear all of the responsibility for how the systems are used, even if they are fully aware of such disparities? Does building and selling a system constitute accepting or endorsing a specific moral and legal interpretation of what constitutes the acceptable use of force?

Consider building existing United States use of force standards and the current practices of police officers into a future automated robocop. Given the recent public outrage over the use of force by police, this ought to be recognized as deeply irresponsible and dangerous. Indeed, as #BlackLivesMatter and CampaignZero have made clear,[10] there is an urgent need to bring the laws and policies of federal, state, and local law enforcement on the use of force into line with international standards.

3. When Is Violent Force and Lethal Force Appropriate, and Against Whom?

This section will examine the international legal standards for the use of force by police, as well as the guidelines issued by United Nations Human Rights Council (UNHRC),[11] ICRC,[12] and Amnesty International,[13] and the legal

[10] CampaignZero.org

[11] "Basic principles on the use of force and firearms by law enforcement officials," Adopted by the Eighth United Nations Congress on the Prevention of Crime and the Treatment of Offenders, Havana, Cuba, 27 August to 7 September 1990, http://www.ohchr.org/EN/ProfessionalInterest/Pages/UseOfForceAndFirearms.aspx

implications of designing robotic systems to use violent and lethal force autonomously. Existing legal standards rely heavily on human judgments, which would be difficult to replicate in a technical system. These judgments require establishing many socially-coded expectations about an individual, their capacity to harm others or themselves, and their intention to do harm to themselves or others. This becomes clear as we start to analyze the actual guidelines that are in place.

Much like the design of autonomous weapons for military use, the use of force by police officers in law enforcement must meet a number of specific conditions in order to be lawful: 1) it must be necessary to prevent an imminent grave bodily harm or the death of a person; 2) it must be applied discriminately, 3) it must be applied proportionately; and 4) the use of force must be accountable to the public. In particular, the UNHRC guidelines state:

> **General provisions**
>
> 1. Governments and law enforcement agencies shall adopt and implement rules and regulations on the use of force and firearms against persons by law enforcement officials. In developing such rules and regulations, Governments and law enforcement agencies shall keep the ethical issues associated with the use of force and firearms constantly under review.
>
> 2. Governments and law enforcement agencies should develop a range of means as broad as possible and equip law enforcement officials with various types of weapons and ammunition that would allow for a differentiated use of force and firearms. These should include the development of non-lethal incapacitating weapons for use in appropriate situations, with a view to increasingly restraining the application of means capable of causing death or injury to persons. For the same purpose, it should also be possible for law enforcement officials to be equipped with self-defensive equipment such as shields, helmets, bullet-proof vests and bullet-proof means of transportation, in order to decrease the need to use weapons of any kind.
>
> 3. The development and deployment of non-lethal incapacitating weapons should be carefully evaluated in order to minimize the risk of endangering uninvolved persons, and the use of such weapons should be carefully controlled.
>
> 4. Law enforcement officials, in carrying out their duty, shall, as far as possible, apply non-violent means before resorting to the use of force and firearms. They may use force and firearms only if other means remain ineffective or without any promise of achieving the intended result.
>
> 5. Whenever the lawful use of force and firearms is unavoidable, law enforcement officials shall:
>
> (a) Exercise restraint in such use and act in proportion to the seriousness of the offence and the legitimate objective to be achieved;
>
> (b) Minimize damage and injury, and respect and preserve human life;
>
> (c) Ensure that assistance and medical aid are rendered to any injured or affected persons at the earliest possible moment;
>
> (d) Ensure that relatives or close friends of the injured or affected person are notified at the earliest possible moment.
>
> 6. Where injury or death is caused by the use of force and firearms by law enforcement officials, they shall report the incident promptly to their superiors, in accordance with principle 22.
>
> 7. Governments shall ensure that arbitrary or abusive use of force and firearms by law enforcement officials is punished as a criminal offence under their law.

[12] International Committee of the Red Cross (2015). The use of force in law enforcement operations, Legal factsheet, September 23, 2015, https://www.icrc.org/en/document/use-force-law-enforcement-operations; International Committee of the Red Cross (2011). Violence and the use of force. Reference Manual, https://www.icrc.org/eng/assets/files/other/icrc_002_0943.pdf

[13] Amnesty International (2015). Use of force: Guidelines for implementation of the UN Basic Principles on the Use of Force and Firearms by Law Enforcement Officials. September 7, 2015, http://www.amnesty.nl/nieuwsportaal/rapport/use-force-guidelines-implementation-un-basic-principles-use-force-and-firearms

> 8. Exceptional circumstances such as internal political instability or any other public emergency may not be invoked to justify any departure from these basic principles.[14]

Following the conclusions of the previous section, what would be required to design a law enforcement robot that conformed to the international standards established by the UNHRC and treaties? That is, as an HRI designer, how should we go about designing the interactions between a robot and the citizens it encounters? Given that the use of violent force and lethal force is only appropriate when there is an imminent threat of severe harm or death to a person, how do we design a system that can recognize threats? What is the legal definition of a threat, what are the conditions for meeting it, how could a system be designed to recognize it, and how can the system correctly identify the agent posing the threat?

3.1. How to Recognize Threats?

The #BlackLivesMatter movement has gained momentum following a series of highly publicized killings of unarmed people of color by police officers, many of which were captured on video by CCTV, police dash-cams, and witness cellphones, which later went viral on social media.[15] In many of these cases, particularly those captured on camera, the individuals who are killed by police do not appear to be acting in the ways described in official police reports, do not appear to be threatening or dangerous, and sometimes even appear to be cooperating with police, attempting to follow police orders, or gesturing at surrender by raising their hands (inspiring the slogan "Hands Up, Don't Shoot!"). As an HRI designer, what types of gestures, actions, and behaviors should count as "threats," or "willingness to cooperate," and how can they be recognized?

Upon seeing the viral videos of violent police encounters, it is quite natural to attempt to "read" these scenes and judge the actions of the suspect and the officer, and to try determining for ourselves whether the use of violence was necessary and appropriate. Of course, the views of the public are not always in line with the perspectives of law enforcement officers and prosecutors. Much of this disparity lies in the professional training of police and the deficient legal standards used by prosecutors in most cases. As an HRI designer, it will be necessary to choose among such perspectives when making design choices.

It is also legitimate to ask why there should be such a disparity between what gestures, actions, and behaviors the public understands as a "threat," compared to what professional law enforcement and experts would recognize as a "threat." One might wish to acknowledge that the professionals have a certain expertise in making such judgments and may believe that this comes from training and experience. However, if one wishes to capture the ways in which the public actually interacts with police officers, it might make more sense to evaluate threats according to the lay perspective that is common within the public. That is, if police are meant to communicate effectively with the public, it would be dangerous for them to have a different understanding and expectation of which gestures, actions, and behaviors constitute a threat than the members of the public do. Otherwise, how are members of the public supposed to know when they are inadvertently making a threatening gesture or how to properly communicate a willingness to cooperate?

There has been much written on how police read and respond to "furtive" movements, and individuals reaching into their pockets, where they might have a weapon. In reality, these judgments are quite subjective and depend heavily on situational context, in which the police officer might be expecting a threat based on the general appearance and manner of an individual. These types of general impressions, which instead could be thought of as prejudice or profiling, can powerfully shape the perception of any actions or utterances by a suspect. In the legal review of such judgments, the legal standard is whether a "reasonable person" in the same situation would have recognized the actions of the suspect as posing a threat. Unfortunately, there is no shortage of experts ready to testify that the simplest of gestures, or even complying with police orders to present identification by reaching into a pocket, could indicate reaching for a weapon, and thus pose a threat.

Indeed, when the video of the 1991 beating of Rodney King by Los Angeles Police Department was subject to expert analysis during the trial, it was deconstructed frame-by-frame to confirm the police report that King posed a

[14] In the work cited by note 12 above.

[15] These include Michael Brown in Ferguson, Missouri; John Crawford III in Beavercreek, Ohio; Eric Garner in Staten Island, New York; Freddie Grey in Baltimore, Maryland; Walter L. Scott in North Charleston, South Carolina; 12-year old Tamir Rice in Cleveland, Ohio; Laquan McDonald in Chicago; and many others.

threat to the 16 officers who were beating and tasering him as he lay face down on the ground, because his ankle moved when he was stuck, indicating an intent to get up and fight back.[16] Of course, this reading of Mr. King's gestures depends on seeing them as gestures, rather than normal reactions to being violently struck, as well as a contextualizing assumption that Mr. King was high on powerful drugs and possessed an almost super-human strength and tolerance for pain. The officers who initially stopped Mr. King claimed that his manner and glazed look indicated to them that he was under the influence of powerful drugs, as did his erratic driving manner.

We should hope that any police robot would do better than the LAPD with regard to the use of force. But it is important to keep in mind that some theory of human gestures, and how they might signify a threat or a willingness to cooperate must be established and built into the HRI design of a law enforcement robot. Which such theories and models should be used? Those devised by the defense "experts" for the police who beat Rodney King? Some other experts who are trained to see furtive movements? Should we train a machine learning algorithm, such as Google DeepMind, to recognize such gestures? Should we try to empirically determine how the community in which the robot will be used would "read" such gestures? Additionally, there could be social and cultural specificity to such gestures, as well as the local laws governing the carrying of weapons, whether it is Sikhs carrying religious knives in India, Pashtun shepherds carrying rifles in Afghanistan, or suburbanites exercising their open-carry rights in the United States.

Furthermore, how might automated technologies be designed to make special considerations for particularly vulnerable populations? There are considerable challenges for police to recognize not only people who may be intoxicated by alcohol or a host of mind-altering drugs, but also for recognizing individuals in need of special consideration during police encounters. Many citizens may not respond to police officers, or police robots, in the manner we might typically expect of a healthy adult. For instance, special considerations ought to be made for the elderly, children, pregnant women, people experiencing health emergencies (including seizures and panic attacks), the mentally ill, and the physically disabled including the deaf, the blind, and those using wheelchairs, canes, prosthetics, and other medical aides and devices. Ultimately, this raises questions about whether automated systems are capable of meeting the legal requirements for the use of force at all.

Designers must make assumptions about the people who may use their technologies. In most cases, they assume that people will fall within the bounds of "normal" in a broad range of ways. Relatively few technological devices are designed to accommodate individuals with special needs. Because of their public nature, many buildings and transit infrastructures are design for accessibility, primarily because they are required to be by law in the United States, and now internationally.[17] Presumably these laws would also require law enforcement robots to recognize the special needs of people with permanent disabilities. It may also require accommodations for individuals who are clearly suffering from temporary episodes or impairments that cause them to behave unpredictably or even pose a threat.

Of course, it is difficult for a human officer to recognize when a suspect is on drugs or suffering a delusional episode. But more effort needs to go into training officers to recognize and deal with common forms of mental illness without the use of force. Several recent cases of police shootings have involved individuals with known mental health issues being shot even when the police were informed of their mental conditions when called to give assistance.[18] In some cases, the deaf are shot for failing to follow verbal police orders, or the blind or disabled are subject to police violence because their physical aides are seen as weapons.[19]

---

[16] Goodwin, Charles (1994). Professional vision. *American Anthropologist, 96*(3), pp. 606-633.

[17] Americans with Disabilities Act of 1990, https://www.eeoc.gov/eeoc/history/35th/1990s/ada.html, and the Convention on Rights of Person with Disabilities, http://www.un.org/disabilities/convention/conventionfull.shtml

[18] Crepeau, Megan, Jeremy Gorner and Grace Wong (2015). "2 fatally shot, 1 accidentally, by Chicago police on West Side; families demand answers," *Chicago Tribune*, December 26, 2015, http://www.chicagotribune.com/ct-chicago-police-shooting-20151226-story.html; Mather, Kate and James Queally (2016). "More than a third of people shot by L.A. police last year were mentally ill, LAPD report finds," *Los Angeles Times*, March 1, 2016, http://www.latimes.com/local/lanow/la-me-ln-lapd-use-of-force-report-20160301-story.html

[19] Jauregui, Andreas (2014). "Deaf man fatally shot by Florida deputy after allegedly 'brandishing firearm'," *Huffington Post*, September 23, 2014, http://www.huffingtonpost.com/2014/09/23/edward-miller-deaf-man-fatally-shot_n_5868538.html; Heath, Brad (2014). "Policeman who shot, killed Detroit man shares his story," *USA Today*, August 26, 2014, http://www.usatoday.com/story/news/nation/2014/08/26/krupinski-detroit-police-shooting/14634913; Carter, Helen (2012).

Another key aspect of detecting a "threat" is to recognize a weapon. A number of recent police shootings have involved toy guns. While it might seem easy to train up a neural network to recognize guns, such an algorithm will not likely be any better than humans at distinguishing toy guns from real guns; though toy guns are required to have bright orange tips, these can be removed or obscured. Indeed, context is important, but several police shootings have occurred on playgrounds[20] and even the toy-section of a Wal-Mart,[21] where one would hope that the default assumption would be that a gun was a toy. There are also situations in which guns are disguised as banal objects.[22]

More problematically, almost any object could conceivably be used as a weapon, though not all with the same degree of threat. A stick or a hammer can be an effective weapon, though they have clear limitations. The level of threat such objects pose as weapons is still much less than a loaded gun, however, and this will be discussed below in the context of proportionality. There are also questions of how robots might interpret citizens who use crutches, canes, walkers, wheelchairs, oxygen tanks, prosthetics, service animals, and other medical aides. These could be used as weapons, but that does not imply that such individuals are always "armed with deadly weapons" and thereby pose a threat. An HRI system would need to be able to recognize such medical aides and accommodate the individuals who depend on them accordingly.

Most banal objects can potentially be weapons, though are only rarely ever used as such. How do we design a system that recognizes them as weapons only when they are being used as weapons? This will be an incredibly difficult technological challenge. It requires not merely object recognition, but understanding both the physical-causal system in which an object can become a weapon and cause physical harm, as well as the psychological intention of an individual to do harm. Recognizing either of these will be extremely difficult technologically, yet absolutely necessary for the lawful use of force.

3.2. Physical Threats Must Be Recognized

Distinguishing when a bodily motion constitutes a meaningful gesture in HRI has primarily focused on clearly established gestures or on training people to perform specific control gestures (e.g., Xbox Kinect, Nintendo Wii, or Leap interfaces). Recognizing "threats" cannot be expected to necessarily conform to trained or pre-existing cultural gestures. Picking out which bodily movements are actually intentional threats requires understanding the situational context of use, the significance of a movement within an ongoing interaction, and maintaining a psychological model of the agent making the movement. Each of these can be challenging for a human police officer but nearly impossible for current and foreseeable HRI technology.

In many cases, people can communicate their intentions verbally. But while speech recognition has gotten quite good (e.g., Apple's Siri), it is still challenging to distinguish which verbal utterances constitute threats or to distinguish serious threats from jokes or sarcasm. Moreover, a verbal threat may not be considered a threat of grave bodily harm or death unless the person making the threat has plausible and available means for carrying it out. And even then, the threat may not be imminent nor require violent or lethal force to avert. It might be possible to talk someone out of carrying out a threat or thwart their capacity to carry out the threat. Indeed, any law enforcement robot should be required to attempt to avert such threats by all available and feasible means before resorting to the use of violent and lethal force.

"Police Taser blind man mistaking his white stick for a samurai sword," *The Guardian*, October 17, 2012, http://www.theguardian.com/uk/2012/oct/17/police-taser-blind-man-stick; Izadi, Elahe (2015). "Video shows Seattle cop arresting elderly black man using golf club as cane," *The Washington Post*, January 29, 2015, https://www.washingtonpost.com/news/morning-mix/wp/2015/01/29/video-shows-seattle-police-officer-arresting-an-elderly-black-man-carrying-a-golf-club

[20] Casarez, Jean (2015). "Grand jury saw enhanced video of Tamir Rice shooting," *CNN*, December 28, 2015, http://www.cnn.com/videos/justice/2015/12/28/tamir-rice-shooting-grand-jury-saw-enhanced-video-casarez-sot-nr.cnn

[21] Remizowski, Leigh (2014). "Grand jury gets case of Ohio man shot by police in Walmart store," *CNN*, September 22, 2014, http://www.cnn.com/2014/09/22/us/ohio-walmart-death/index.html; Shiochet, Catherine E., and Nick Valencia (2014). "Cops killed man at Walmart, then interrogated girlfriend," *CNN*, December 16, 2014, http://www.cnn.com/2014/12/16/justice/walmart-shooting-john-crawford

[22] Heisler, Yoni (2016). "Controversial new handgun design looks exactly like a smartphone," *BRG*, March 25, 2016, http://bgr.com/2016/03/25/smartphone-gun-ideal-conceal

One advantage that robotic law enforcement will have over human police officers is that they will not be people, and thus will not need to act in self-defense. Indeed, they would have no right to defend themselves with violent and lethal force in virtue of not being persons, and thus not persons who could be threatened with grave bodily harm or death. As objects, they are only threatened with damage, and damage to property, such as the robot, does not warrant intervention by violent or lethal force in most situations. A robot could only intervene with violent or lethal force when a person other than the robot was under threat. In some cases, the person threatened may also be the person posing the threat, (i.e., threats of self-harm and suicide). In such cases, much like interactions with the mentally ill mentioned above, special techniques are called for to defuse the situation. It simply makes no sense to use lethal force against someone who is threatening only themselves. Some lesser violent force might be appropriate, however, in order to avert a greater threat of harm such as suicide. Since most instances of the use of force by police involve threats to the police officer, a robot may be advantageous in dealing with dangerous individuals due to the fact that they need not act out of fear for their own safety. This carries with it a requirement for robots to use much less force than potentially lethal force, and only if there is another person around who is under imminent threat.

Much of the interpretation of verbal and gestural intentions seems open to differing subjective perspectives, yet the law requires an objective standard of interpretation. In Graham v. Connor, the Supreme Court established the legal standards that the use of force is "objectively reasonable in light of the facts and circumstances confronting them" from the perspective of a "reasonable officer on the scene."[23] Of course this standard has been stretched, and perhaps abused. We saw in the previous section that there is no simple way to recognize weapons, nor is there necessarily a clear pattern of interaction that constitutes a threat, such as "failing to follow lawfully issued directions" or a set of standard "suspicious behaviors." The recognition of a threat requires a human-level understanding of the facts and circumstances, as a reasonable human officer might have. It is not clear when or if robots will achieve such capabilities.

Given the difficulty of estimating the intention or determination of a person to inflict severe injury, is it better to assume the worst? Or the best? Or to develop the best possible model of intention given what is known, and thus acting on a model that is known to be uncertain, as long as it is the best available? Or to wait to act only when there is certainty, or a sufficient degree thereof? Should HRI designers be the ones responsible for making these decisions, and setting the certainty parameters? Indeed, in most real-world cases, the police officer makes these discretionary judgments, often with little accountability.[24] It is also not clear how often the human officers get it right or wrong in anticipating threats.

Beyond the fundamental technical and moral issues with machines automatically categorizing human actions and intentions, they must also be able to make complex judgments about causal physical systems in order to appreciate the imminence, likelihood, and severity of the completion of a threat. It is quite conceivable that robots will eventually have algorithms that allow them to simulate and model the physical dynamics of the world, at least in simple ways necessary to interact with physical objects. As such, they may be able to make certain predictions about how physical events might unfold in the future. Insofar as those are well-behaved physical systems, with tractable degrees of complexity and uncertainty, we might expect predictive algorithms to do as well or better than humans in such predictions. This could work only when we understand the causal dynamics of physical systems well enough, and could recognize them in a given system with available sensor data, and model them accurately enough and fast enough to act accordingly (where multiple potential actions must be simulated in order to choose the best). This is only possible today for a few simple systems, such as inverted pendulums, juggling balls, or avoiding stationary obstacles, or constrained environments such as manufacturing automation and self-driving cars. But to fully predict the range of potential threats might also require understanding chemistry and the functioning of complex mechanical and electronic systems, including in non-standard ways (e.g., the effects of dropping a toaster in water when it is plugged in or unplugged).

---

[23] U.S. Supreme Court, (1989). "Graham v. Connor," 490 U.S. 386, argued February 21, 1989, decided May 15, 1989, https://supreme.justia.com/cases/federal/us/490/386/case.html.

[24] Elizabeth E. Joh (2016). New surveillance discretion: Automated suspicion, big data, and policing. *The Harvard Law & Policy Review, 10*, pp. 15-42, http://harvardlpr.com/wp-content/uploads/2016/02/10.1_3_Joh.pdf

It is not implausible that sufficient research efforts into this area will yield increasing capabilities to model and simulate more complex dynamic systems with greater precision, fewer constraints, and that robots will become better at choosing appropriate actions to take in relation to unfolding causal systems. Such insight and understanding of physical systems would bring greater understanding of how to interfere with dynamic systems so as to avert or thwart the threat. Such understanding would necessarily imply a responsibility to direct any actions to do so in a way that did not involve violent or lethal force unless no other option was available, which might turn out to be quite rare. Bullets and blows might be intercepted and blocked, those threatened might be shielded, dangerous forces might be redirected, and potential victims might be moved out of harms' way. Consequently, there would be a responsibility to avoid the use of violent and lethal force, within the capabilities of the robotic system, until all other alternatives are exhausted. Many of these questions regarding the severity of a given threat also relate directly to the question addressed in the next section—what constitutes a proportionate response?

3.3. Threat Requires Intention

Physical systems, while complex, might be subject to reliable simulation and prediction. The same is not necessarily true of predicting human decisions, actions, and intentions. It is well known that social systems, and psychological systems, are not strictly predictable in the same sense as physical systems.[25] The best available quantitative and statistical methods cannot actually predict how any individual person will react to a stimulus, who they will vote for on Election Day, or how they will act in a given situation. Of course, studying individuals and populations to determine the correlates and causes of typical, median, and majority behaviors and social norms, or of behaviors that are atypical, divergent, or deviant from social norms,[26] can provide insights into social systems and the human experience, and could be effective in encouraging or discouraging certain behaviors or influencing individuals through communication and coercion. But such scientific understanding is not, strictly speaking, predictive of individual behaviors in individual situations.

While positivist social scientists have long sought to emulate the precision and predictive powers of the physical sciences, there are fundamental hurdles to doing so. One can argue that this is due to lack of experimental control, imprecise measurement, insufficient conceptual clarity or theoretical understanding, or simply due to human creativity and free will. Economists, for instance, have long understood that attempts to produce "perfect" models of market behavior will inevitably influence the very markets under study, and thus change the very behaviors they are attempting to predict—whether self-fulfilling or self-defeating their own predictions.[27] The same might well be argued for policing interventions, wherein the escalation of force by an officer results in the greater resistance or violent response of a suspect, or where the effort to de-escalate a situation brings the suspect back to an interaction that might have otherwise turned violent.

These reflections on the fundamental causal uncertainty of human actions are not hypothetical, and it would be dangerous to ignore them when considering how to program a robocop. By "locking in" a model of human action into the predictive simulator of a robot, we could, in effect, be instigating the very behaviors that the system is predicting. Even if this only occurs in a low percentage of cases, it should be a concern for policy-makers. Even if big data techniques might give spectacular statistical predictions of the probability that an individual will act a certain way, that is not the same as knowing how they will act, nor is it the same as understanding why they do act a certain way. We might call this the epistemic bounds on predicting human actions and behaviors. In situations where the stakes are high, such as the deprivation of human rights to life or bodily integrity, even the best available predictions may not be sufficient justification for an irrevocable action.

Beyond the epistemic limits of imposing behavioral models on individual choice and actions, there are ethical and moral considerations. In particular, treating individual persons as merely sums of their aggregate features and probabilistic propensities is to treat them as objects and not as moral subjects—as means and not ends in the Kantian sense. We may be able to predict the likelihood of someone purchasing a book on Amazon based on their other purchases, but that does not begin to tell us *why* they purchase that book, or the other things they purchase. Of

[25] Peter Winch, (1958/1990). *The Idea of a Social Science and its Relation to Philosophy, 2nd Edition*. London: Routledge.
[26] Howard S. Becker (1963). *Outsiders: Studies in the sociology of deviance.* New York: The Free Press.
[27] For example, predicting a bank collapse can instigate a run on the banks, while predicting the rise of a stock price can contribute to its price inflation.

course, Amazon need not care about the reasons, as long as they can use those predictions to make more sales. But if we are designing a system with the authority to deprive individuals of their basic human rights, we need to treat them as legal and moral persons. Under the current legal system, individuals are judged by their beliefs and intentions, as well as their overt and objective actions. Perhaps the gravest danger of automating legal and moral decisions is that there is no clear technological means for determining or judging the beliefs and intentions that guide the actions of others.

Similarly, the choices made by police officers on how to respond to threats require psychological skills of interpreting a given situation, assessing the intentions and motives of the people involved, assessing how the individuals involved will interpret and react to the actions taken by the officer, further cascading actions and responses, and weighing the risks of various outcomes against the uncertainty of their own assessment of the situation. Of course, as such situations unfold, the interpretive understanding of the situation, the individuals involved, and their intentions, shift and evolve. As officers gain more information about the situation through questioning and observation, they also develop their understanding of who they are dealing with and how and why they may act or react.

It is important to note here that even in an ideally operating robocop, there is a clear sense in which we dehumanize the citizens who are policed by treating them as objects rather than subjects. This can, for certain technologies, be rectified after the fact through accountability mechanisms. For instance, traffic cameras detecting speeding cars or red-light violations essentially objectify drivers and do not allow them to explain their actions (e.g., speeding a mother in labor to the hospital) as they might to an officer if they were pulled over. They could, however, make such appeals and provide explanations after the fact and appeal the citation. This is not true for irrevocable deprivations of rights. Most clearly in the use of lethal force—no appeal can bring back the dead. This is also true of the violation and loss of bodily integrity and human dignity that comes from other uses of force or deprivations of freedom. Despite the payment of monetary damages or the healing of wounds, the injustice of such violations have irrevocable consequences.

4. How Much Violent and Lethal Force Is Appropriate and Proportional to a Given Threat?

We turn now to a closely related problem—given that the use of violent and lethal force is determined to be appropriate, how much force should be used? Deciding how much force is appropriate in the given circumstances, and when and how to escalate the use of force, is known as *proportionality* in the use of force. Again, there are questions of which legal standards to conform to, but also much more challenging technical issues involving how to meet those requirements given that they demand explicitly *human* judgments.

Based on the previous section, it should be clear enough that even in ideal conditions and situations, it will be incredibly challenging to preprogram a system to determine whether the use of force is appropriate, and similarly to determine what level of violent or lethal force is appropriate. Moreover, if such systems are actually sophisticated enough to model the dynamic physical systems within which threats are framed, then they will likely have insights into means of intervening that do not necessitate the use of violence or lethal force against the individual posing the threat.

Consider someone wielding a blunt weapon and threatening other people with it. A robot might be able to grab the weapon, or put itself between the threatening person and those being threatened to block any blows, or something even more clever, all before it might consider using violent force. Moreover, it need not, and under the international guidelines for the use of force by police *should not*, resort to the use of firearms or lethal force when other means are available for dealing with the threat. Even if a firearm is used, it could be directed at the hand or foot of the threatening individual rather than the head or chest, in order to use the minimum violence necessary to neutralize the threat.[28]

[28] It is thus disconcerting that most police officers in the United States are trained to aim shots for the head or chest in all cases, or "shoot to kill" *by default*. This approach is built on a series of assumptions that if a firearm is being used it must already be the case that there is a threat of death, and targeting the body mass is the most likely way to stop a threatening person. This approach, however, precludes significant proportionality judgments being made once the firearm is drawn. See Visser, Steve (2016). "Why do police shoot to kill?" *CNN*, November 30, 2016, http://www.cnn.com/2016/09/28/us/why-police-shoot-to-kill. Police in

In legal terms, a proportionality judgment is not simply a matter of deciding what action will neutralize a threat with the minimal necessary force. It is also necessary to weigh the nature and severity of the threat against the nature and severity of the violence aimed to neutralize it. These judgments require not only estimations of the probability of various outcomes but the values of those outcomes. In general, it would be disproportionate to shoot someone who is threatening to punch someone—unless it is reasonable to expect the punch to be as damaging as the gunshot. Furthermore, apprehending or incapacitating a person is generally sufficient to thwart threats not already set in motion. Apprehending someone, including arrest, involves the use of force that could be violent, could result in injury, and necessarily deprives an individual of the freedom of movement. While preferable to violent and lethal force, physical restraint must also be weighed against the threat posed, even when it is difficult to quantify and compare such options.

There is a technical and moral issue here regarding whether an artificial system can make the type of value judgments that are constitutive of proportionality judgments in the use of force. I have made the similar arguments with regard to proportionality in the use of lethal force by military robots in armed conflict.[29] In a military context, the proportionality judgment in an attack requires understanding the value of a military objective and weighing that value against the negative value of the risks posed to civilians and civilian infrastructure in a given attack. Something similar is required in police use of force, yet even more must be taken into consideration—including the rights and bodily integrity of the person against whom violence is directed. Such consideration is not required in armed conflict but is required in policing.

This problem is even more severe for the use of force in law enforcement than it is in military armed conflicts, insofar as killing or harming a citizen is never a law enforcement objective in itself. In armed conflict, it can be argued that killing an enemy combatant is itself a military objective. But killing a criminal suspect can never be a law enforcement objective. Protecting people from an imminent threat of death or severe bodily harm is the only law enforcement objective that can justify the use of lethal force, and the use of such force is only a means, not an end. Similarly, a threat to use violence can be just as effective as the actual use of violence in many cases. Thus, merely pointing a weapon and shouting, "Stop! Drop your weapon!" ought to be attempted before using actual force, when feasible. And again, making a feasibility decision, and how much time one has to consider and attempt alternatives to violent force, will be quite complex and probabilistic at best.

Finally, police and potential police robots must also be capable of recognizing opportunities for de-escalating a threat. If a suspect throws up their hands and says, "Don't shoot!" or makes similar symbolic acts to that effect, the robot must also de-escalate its use of force. Of course, such a robot might get fooled, but it has to provide that opportunity to all suspects. It is tempting as engineers to think that we might provide a sophisticated model of risk assessment and decision theory for proportionality judgments. But it is clear in the law that a human must make such decisions, both because such technological solutions are as yet inconceivable, but also because that entails a human who is responsible and accountable for the use of force. This brings us to a final question.

5. Who Will Be Responsible for the Violence a Robot Commits, and How Will They Be Accountable?

Like law enforcement officers, a law enforcement robot system must be accountable for any use of force. At the very least, this would entail transparency with regard to its algorithms and functioning, as well as logs of its operations and black box recordings of its interactions. The public should have a right to know how these systems operate, and the basis upon which they might decide to use force. Even if some may use this information to "game the system" or exploit weakness in the robot, the social costs of deploying violent police robots with opaque reasoning behind their uses of force would be far worse.

Europe and other countries are trained instead to aim for legs and feet by default. See Stute, Dennis (2014). "Why German police officers rarely reach for their guns," *Deutsche Welle*, August 27, 2014, http://www.dw.com/en/why-german-police-officers-rarely-reach-for-their-guns/a-17884779

[29] Asaro, Peter (2012). On banning autonomous lethal systems: Human rights, automation and the dehumanizing of lethal decision-making. Special issue on new technologies and warfare. *International Review of the Red Cross, 94* (886), Summer 2012, pp. 687-709, http://www.icrc.org/eng/resources/international-review/review-886-new-technologies-warfare/index.jsp

Individual police officers must be accountable for their actions to superiors, but also to the communities which they serve. This has led to calls for community review boards for police conduct. We might also consider community review boards for robots. The challenges facing such boards would include their ability to assess the technological capabilities of robotics systems, especially with algorithms that are complex, context dependent, and perhaps not transparent. Such review boards could also analyze data for the deployment of robots and logs of their interactions with members of the public. Any systemic flaws in the functioning a law enforcement robot, or systemically unfair deployments, ought to be auditable, and all complaints should be investigated and adjudicated where necessary.

Even with such transparency, we cannot really hold robots legally responsible for their actions. Further, it is awkward or impossible to hold programmers responsible. Product liability might come into play if there were negligence on the part of manufacturers or designs known to be faulty were released, but in most cases, reasonability standards would permit a broad range of potentially harmful effects. This is a good reason for HRI designers and roboticists to consider a code of ethics that precludes the use of violent and lethal force by robots altogether. Police departments might well be held liable in lawsuits over specific cases of the use of force by its robots. This might ensure that particular robots are kept up in proper maintenance and software updates. It would prove more difficult to hold police departments liable for civil rights violations if those robots perform in systematically racist or otherwise discriminatory ways.

Another option would be to design systems to make specific humans responsible for their actions. Various approaches have been proposed for lethal military robots, including designing explicit lines of human responsibility into systems,[30] turning humans into "moral crumple zones" in order to absorb responsibility (though not necessarily deserving it),[31] and requiring meaningful human control over the use of violent and lethal force.[32] As in military systems, there are many reasons to retain human control over the use of violent and lethal force in policing robots. Accountability for the use of force is far more critical in civilian policing than in armed military conflicts, both legally and politically. Not only should police robots meet whatever standards emerge from international discussion of the autonomous weapons in armed conflict, there should be even stricter standards for human control and responsibility than exist for military systems.

6. Conclusions—Call for an HRI Code of Ethics

It is already understood that robotic systems pose serious dangers to humans. Indeed, it is only recently that robotic systems have been rendered safe enough to work together closely with humans in a broad range of co-robotics applications.[33] Thus far, the history of managing the harms that robots might do to humans has been to reduce the risk of harms wherever possible. This would likely have pleased Isaac Asimov, whose 1st Law of Robotics stated that "A robot may not injure a human being or, through inaction, allow a human being to come to harm." There are various problems with Asimov's Laws as a basis for robot ethics, but this provides a good point of departure for considering the problem of designing systems to use violent and lethal force against humans. That is to say *all* such systems violate the 1st Law of Robotics insofar as they deliberately deploy violence to cause injury to people. From a design perspective, this is fundamentally different than designing a system to minimize harms from actions and activities that are not intended to cause injuries—even if it is known that there are risks of the system failing and thus some probability that it will cause injuries.

[30] Gert-Jan Lokhorst and Jeroen van den Hoven (2011). Responsibility for military robots. In *Robot Ethics: The Ethical and social Impacts of Robotics*. Cambridge, MA: MIT Press, pp. 145-156.

[31] M. C. Elish (2016). Moral crumple zones: Cautionary tales in human-robot interaction. March 20, 2016, We Robot 2016. Working paper, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2757236

[32] Human Rights Watch (2015). Killer robots and the concept of meaningful human control. Memorandum to the Convention on Conventional Weapons (CCW) Delegates, https://www.hrw.org/news/2016/04/11/killer-robots-and-concept-meaningful-human-control

[33] Frank Tobe (2015). Why co-bots will be a huge innovation and growth driver for robotics industry. *IEEE Spectrum*, December 30, 2015, http://spectrum.ieee.org/automaton/robotics/industrial-robots/collaborative-robots-innovation-growth-driver

I conclude that it makes sense to draw a clear line here, and for HRI researchers to refuse to design such systems on ethical and moral grounds. The consideration of a police robot has demonstrated some of the reasons why designing such systems is fraught with perils and challenges that undermine our hopes for the possible benefits of such a system. While these can be framed as technological issues to be sorted out through future research, each of the sections disclosed legal and moral issues that are not addressable through better engineering.

HRI researchers have an ethical duty to consider the social, ethical, and legal context in which the systems they develop will operate. In the case of automating the use of violent and lethal force by police, it is necessary to examine the social, cultural, political and economic contexts in which such systems will operate, as well as the legal and ethical frameworks in which robotic systems may act. This means recognizing the significance of making design decisions for an application area that has significant social implications, but also requires engaging various perspectives on the problems.

The choice of standards to meet is itself an ethical question. Simply adopting the existing legal standards in the United States would be ethically problematic at best, given the degree to which they fall far short of international legal standards and vary greatly between jurisdictions. Building such standards into a HRI system could amount to enabling and perpetuating serious deprivations of human rights. It would be unethical to develop systems that fail to meet international standards of the use of force by police. The fact that current standards in the United States fall below international standards is no excuse for designers and engineers to perpetuate or endorse the flagrant violation of human rights those flawed standards enable.

In considering whether, or how, to automate decisions to use violent and lethal force according to the international standards, there remain a number of significant ethical challenges. While engineers and designers may be eager to operationalize abstract legal concepts and terms into forms that can be more clearly implemented, it is necessary to consider whether such reinterpretations are legitimate. This kind of operationalization is a form of translation, in which an abstract concept is translated into a set of observable concrete features. While this can be an effective means of practical problem solving, it can also result in obscuring or eliminating essential aspects of a concept. This is especially true of many humanistic and psychological concepts embedded in legal standards. Translating "threat" into sets of observable behaviors or motions divorces it from the situational and contextual meaning it had. This is true of both the physical and psychological interpretation of threats.

To the extent that law enforcement robotics can develop the sophisticated HRI that would be required to recognize threats, and the causal systems in which they operate, there is a duty for robotics engineers to devise new means for neutralizing threats of grave harm and death without resorting to the use of violent or lethal force by robots. While this is an added requirement and burden that human law enforcement officers are rarely held to, the moral engineer ought still to strive for it. The ideal for the engineer should be the law enforcement system that can serve and protect everyone in the community, even while it de-escalates, diffuses, and thwarts threats of all kinds, including those from malicious people.

It is thus important to continue to limit the use of violent and lethal force to humans who are properly trained, and who operate in accordance with international standards, and who are accountable to superiors and the communities they serve. HRI researchers should uphold this principle by refusing to develop any systems that deploy violent and lethal force against humans. This might initially take the form of a moratorium, while HRI researchers develop a code of ethics and work for further professional and legal standards that similarly require human judgment in the use of violent and lethal force.

Acknowledgements

This research was supported as part of the Future of Life Institute (futureoflife.org) FLI-RFP-AI1 program, grant #2015-144580. I would like to thank the Center for Information Technology Policy at Princeton University for the resident fellowship during which this paper was written, and The New School for the sabbatical leave that made this possible.

P. Asaro, School of Media Studies, The New School, NY, USA; Center for Information Technology Policy, Princeton University, NJ, USA; Center for Internet and Society, Stanford Law School, CA, USA. Email: asarop@newschool.edu