*****PRELIMINARY DRAFT***3-20-16***DO NOT CITE******

Will #BlackLivesMatter to Robocop?[1]

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

Introduction

[1] In keeping with Betteridge's law of headlines, one could simply answer "no." But investigating why this is the case is still worthwhile.

#BlackLivesMatter is a Twitter hashtag and grassroots political movement that challenges the institutional structures surrounding the legitimacy of the application of state-sanction violence against people of color, and seeks just accountability from the individuals who exercise that violence. It has also challenged the institutional racism manifest in housing, schooling and the prison-industrial complex. It was started by the black activists Alicia Garza, Patrisse Cullors, and Opal Tometi, following the acquittal of the vigilante George Zimmerman in the fatal shooting of Trayvon Martin in 2013.[2]

The movement gained momentum following a series of highly publicized killings of blacks by police officers, many of which were captured on video from CCTV, police dashcams, and witness cellphones which later went viral on social media. #BlackLivesMatter has organized numerous marches, demonstrations, and direct actions of civil disobedience in response to the police killings of people of color.[3] 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 or trying to follow police orders.

While the #BlackLivesMatter movement aims to address a broad range of racial justice issues, it has been most successful at drawing attention to the disproportionate use of violent and lethal force by police against people of color.[4] The sense of "disproportionate use" includes both the excessive amounts of forced used in a given encounter, and the frequency with which force is used in police encounters with people of color. Since the movement began, a number of journalists, organizations and institutions have produced studies and reports investigating both racism in policing and the use of force by police.[5] Collectively, these raise a series of questions about the legitimate use of violent and lethal force by police, and the legal regulation of inappropriate and unnecessary use of force by police.

Due to the increased media attention given to police violence when a video of the incident is available, many people have called for requiring police to wear body-cams to record their

---

[2] https://en.wikipedia.org/wiki/Black_Lives_Matter

[3] 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.

[4] The #AllLivesMatter hashtag appears to be largely aimed at diffusing or rejecting the racial critique presented by #BlackLivesMatter. This paper does not endorse that political reaction or its aims, but will consider the implications of automating police use of force on all citizens as well as its disproportionate effects on particular racial and disenfranchised groups.

[5] These include civil rights investigations by the Department of Justice of the Ferguson, Missouri PD (http://www.justice.gov/sites/default/files/opa/press-releases/attachments/2015/03/04/ferguson_police_department_report.pdf , https://www.washingtonpost.com/news/post-nation/wp/2015/03/04/the-12-key-highlights-from-the-dojs-scathing-ferguson-report/), and the Albuquerque, New Mexico PD (http://www.justice.gov/sites/default/files/crt/legacy/2014/04/10/apd_findings_4-10-14.pdf), Amnesty International's report on police use of force and firearms in the US (http://www.amnestyusa.org/research/reports/deadly-force-police-use-of-lethal-force-in-the-united-states), and numerous journalistic investigations into a range of topics including inadequate training of police to deal with the mentally ill (http://www.washingtonpost.com/sf/investigative/2015/06/30/distraught-people-deadly-results/)

interactions with the public (cite). While this appears to be a potential technological solution to a set of problems, it has obvious limitations. In particular, a number of the high-profile cases did involve police body-cams as well as police car dash-cams, and yet many of the same accountability problems persist–police are not charged, indicted or convicted despite the videos. The public discussion of police body-cams points to both a widespread desire for a simple technological solution to complex social problems, and an awareness of the potential power of surveillance on accountability, even as it fails to address the social and legal frameworks within which these technologies function.

As a means of critiquing this discussion of body-cams and other policing technologies as solutions to the social problems manifest in policing, this paper will consider an even more sophisticated policing technology: a hypothetical RoboCop. That is, if we wished to address the various forms of racism, psychological aggression and abuses of power by automating the work of police and particularly the use of force by police, could this work, and if so, would it be desirable?

Recently there have been a number of robotic systems introduced for law enforcement, security and policing.[6] Some of these robots feature weapons such as tasers and tear gas which could be used against people.[7] Additionally, there is growing use of face-recognition[8] and automatic license-plate readers by law enforcement agencies.[9] Admittedly, the RoboCop depicted in the Hollywood sci-fi movies was actually a human police officer whose brain is grafted into a robotic body. For the purposes of this paper I will examine the possible future application of robotics to policing with the understanding that these will be systems that are controlled by programmed computers, rather than cyborgs.[10] In particular, this paper will examine the legal and moral requirements for the use of force by police, and whether robotic systems of the foreseeable future could meet these requirements, or

---

[6]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/).

[7]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-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).

[8]Kelly Gates (2011) *Our Biometric Future*, NYU Press.

[9]https://www.eff.org/deeplinks/2015/10/license-plate-readers-exposed-how-public-safety-agencies-responded-massive

[10]Though it is worth noting that both in the original 1987 film and its recent 2014 remake, the human element is included in order to legitimize the automation of policing and its use of force. The ED-209 was, by contrast, an autonomous lethal military weapon system deemed too dangerous for civilian law enforcement.

whether those laws may need to be revised in light of robotic technologies, as some have argued.[11]

Beyond this, I will consider the racial dimensions of the use of force by police, and how such automation might impact the discriminatory nature of police violence. Many people are inclined to believe that technologies are politically neutral, and might expect a future RoboCop to be similarly neutral, and consequently expect it to be free from racial prejudice and bias. In this way, RoboCop might be seen by some as a technological solution to racist policing. However, many scholars have argued that technologies embody the values of the society that produces them, and often amplify the power disparities and biases of that society. In this way, RoboCop might be seen as an even more powerful, dangerous and unaccountable embodiment of racist policing.[12]

The paper will proceed by examining the problems of racist policing from a number of diverse perspectives. This will include examining the national and international legal standards for the use of force by police, as well as the guidelines issued by UN Human Rights Council,[13] ICRC,[14] and Amnesty International,[15] and the legal implications of designing robotic systems to use violent and lethal force autonomously.

From another perspective, the paper will consider the ways in which digital technologies are not racially neutral, but can actually embody forms of racism by design, both intentionally and unintentionally. This includes simple forms such as automatic faucets which fail to recognize dark skinned hands,[16] the intentional tuning of color film stock to give greater dynamic range to white faces at the expense of black faces,[17] and the numerous challenges of applying facial recognition technologies to racially diverse faces.[18] In other words, how might automated technologies that are intended to treat everyone equally, fail to do so? And further, how might automated technologies be expected to make special considerations for particularly vulnerable populations? The paper will also consider the challenges of recognizing individuals in need of special consideration during police

---

[11] UN Special Rapporteur for Extrajudicial Executions, Christof Heyns, has argued that armed police robots would necessitate new rules for the use of force:
http://www.ohchr.org/en/NewsEvents/Pages/DisplayNews.aspx?NewsID=14700&LangID=E
http://www.ohchr.org/EN/HRBodies/HRC/RegularSessions/Session26/Pages/ListReports.aspx
Amnesty International has also called for banning armed robots in policing:
https://www.amnesty.org/en/latest/news/2015/04/ban-killer-robots-before-their-use-in-policing-puts-lives-at-risk/

[12] This view is captured elegantly in the satirical headline: "New Law Enforcement Robot Wields Excessive Force of Five Human Officers," *The Onion*, June 5, 2014, VOL 50 ISSUE 22. (http://www.theonion.com/article/new-law-enforcement-robot-can-wield-excessive-forc-36220?)

[13] http://www.ohchr.org/EN/ProfessionalInterest/Pages/UseOfForceAndFirearms.aspx

[14] https://www.icrc.org/en/document/use-force-law-enforcement-operations
https://www.icrc.org/eng/assets/files/other/icrc_002_0943.pdf

[15] http://www.amnesty.nl/nieuwsportaal/rapport/use-force-guidelines-implementation-un-basic-principles-use-force-and-firearms

[16] http://mic.com/articles/124899/the-reason-this-racist-soap-dispenser-doesn-t-work-on-black-skin

[17] http://www.vox.com/2015/9/18/9348821/photography-race-bias,
http://www.buzzfeed.com/syreetamcfadden/teaching-the-camera-to-see-my-skin#.ln77Xb361

[18] http://gizmodo.com/5431190/hp-face-tracking-webcams-dont-recognize-black-people
http://mic.com/articles/121555/google-photos-misidentifies-african-americans-as-gorillas

encounters, such as the elderly, children, pregnant women, people experiencing health emergencies, the mentally ill, and the physically disabled including the deaf, blind and those utilizing wheelchairs, canes, prosthetics and other medical aides and devices.

The paper will consider the systemic nature of racism. The automation of policing might fail to address systemic racism, even if it could be successful in eliminating racial bias in individual police encounters. In particular, it will consider the likely applications of data-driven policing. Given the efficiency aims of automation, it seems likely that automated patrols would be shaped by data from previous police calls and encounters. As is already the case with human policing, robotic police will likely be deployed more heavily in the communities of racial minorities, and the poor and disenfranchised where they will generate more interactions, more arrests, and thus provide data to further justify greater robotic police presence in those communities. That is, automated policing could easily reproduce the racist effects of existing practices and its explicit and implicit forms of racism.

Finally, the paper will reflect on the need for greater community involvement in establishing police use-of-force standards, as well as the enforcement of those standards, and other norms governing policing. Moreover, as policing becomes increasingly automated, through both data-driven and robotic technologies, it is increasingly important to involve communities in the design and adoption of technologies used to keep the peace in those communities. Failing to do so will only further increase an adversarial stance between communities and their police force.

The problem of racist policing has multiple causes, and eliminating the problem will require numerous policy, as well as social, changes. I believe it is worthwhile to consider what it would mean to create an automated robotic police officer, and what it would require to ensure that it was not racist, in order to better understand the challenges of eliminating racist police practices in human police officers. In particular, I do not want to suggest that such a technology would be a solution to the problem of racist policing. Indeed, I will argue that there can be no easy technological fix to this problem. Moreover, I want examine the legal, psychological and moral complexity involved in decisions by police officers to use violent and lethal force both as a means to argue against any proposal to authorize automated systems to use violent and lethal force against people, and to further inform and enlighten the current discussions of the use of violent and lethal force by police.

What is meant here by conjuring the notion of a robocop is not exactly what is depicted in the Hollywood films produced in 1987 and 2014. It is far to easy to say “Imagine a robot that perfectly applied the established standards for the use of force, and did so without regard to bias or prejudice, racial or otherwise.” Such an ideal fantasy might be seductive when viewed from a distance, but viewed up close, from the perspective of an engineer who might wish to design such a system, there are deep philosophical and legal issues that make this ideal infeasible, undesirable, and dangerous. Many of the same issues confront other technologies which might be offered as easy technological fixes for the problem of racist policing, such as requiring police to wear body-cams.

In order to automate the use of violent and lethal force in our hypothetical robot, we must start by considering what standards ought to be implemented by our system. This is perhaps the most

significant challenge facing both the elimination of racist policing in the United States, and the hypothetical automation of police use of force. In the first section I examine the international standards for the use of violent and lethal force by police. This will include both the technical challenges, or impossibility, or designing a system that could meet existing international standards, as well as the reviewing the ways in which existing policies within the United States, including federal, state and local laws, all currently fail to meet international standards.

The basic challenges of automating the use of force apply in all situations, regardless of racial context. There are, however, ways in which racism can be embedded in technologies themselves. The second section will examine several examples of automation technologies which manifest racial discrimination. Racial discrimination can be embeded in technology in numerous ways, whether intentionally or unintentionally. This section will review the substantial literature on racialized technologies, and how these might be realized in a hypothetical robocop. While we might hope that the technologies we build will be free from racial bias and discrimination, freeing technologies from such biases will actually require careful and conscious design choices to identify and eliminate that racism at every level of design.

While racism is most recognizable in its overt and egregious manifestations, it also exists within persistent and systemic forms that are much more difficult to recognize, challenge and eliminate. In the fourth section of this paper, I will consider how even a robocop that followed use of force guidelines perfectly, and was completely free of any embedded racism, could still be used to enact and replicate systemic racism.

Perhaps the most significant policy challenge facing the elimination of racist policing, and the excessive use of violent and lethal force by police more generally, is the lack of accountability for police use of force. Fixing the accountability problem for policing in the United States will require significant policy changes. And again, there is no clear or simple technological solution to this problem. Indeed, the introduction of technologies such as body-cams, or even an automated robo-cop, can just as easily serve to justify failures to hold police accountable or further obscure accountability by adding new layers of opacity and new challenges for holding individual officers and police departments accountable for the use of violent and lethal force against citizens.

And finally, I conclude with a summary of the most critical issues facing the reform of standards for the use of violent and lethal force by police, the automation of the use of violent and lethal force by machines, and the overarching necessity for reliable systems of accountability at multiple levels.

Part I: Automating Standards for the Use of Lethal and Violent Force by Police

The most conspicuous manifestation of racist policing is the excessive use of force and lethal force against people of color. The causes of this problem are many and complicated. Indeed, #BLM affiliated Campaign Zero calls for a significant number of policy changes to address this problem.

Their Policy Agenda[19] calls for 30 specific areas in need of legislative and policy reform, at the federal, state and local levels. These areas are categorized under the headings of: Interventions that target racial profiling, broken-windows policing, and for-profit policing; Interactions that target use of force standards including using the least amount of force necessary and restricting lethal force to imminent threats only, providing necessary training for use of force and racial bias, de-militarization of police forces, and promoting diversity in police hiring; Accountability for police through mandated body-cams, civilian oversight of police misconduct and discipline, independent investigators for police killings, lower standards of proof for civil cases against police, revising police contracts that inhibit investigations and civilian oversight of police conduct.

The notion of designing and deploying a robocop that could use violent and lethal force against citizens is fraught with moral and social issues. This paper will consider the hypothetical development of such a system primarily as a foil to reveal the depth and seriousness of these issues, many of which are social rather than technical in nature. My overwhelming concern is to disarm the view that such a system would automatically, necessarily, or by definition, be free from legitimate criticisms of racial bias. To the contrary, it would be easy to intentionally design a robocop to be racist, and quite difficult to design one that is not, given the existing standards, norms, and policing strategies.

Among the various activities the police typically perform, the most morally and politically significant involve the use of violent and lethal force against citizens. Accordingly, the most challenging issue facing the design of our hypothetical robocop will be how to design the algorithms that control the decisions to use lethal and violent force. In technological terms, it is already possible to design a system that is capable of targeting and firing a lethal weapon, such as a gun with some degree of accuracy. Far more challenging is to design a system that only uses force when it is necessary, from a legal perspective, which uses that force discriminately, and to use that force proportionately. Beyond the technical challenges of building a system that can adhere to given rules for the necessity of the use of force, discrimination and proportionality, there are also serious questions about which rules ought to be adhered to, or "built in" to the system, and how those rules ought to be interpreted in actual situations.

Existing standards rely heavily on human judgements, which would be difficult to replicate in a technical system. This requires establishing many socially-coded expectations about an individual, their capacity to harm to others or themselves, and their intention to do harm to themselves or others. Supposing such a system were developed, there is a serious question about how to assess it. If it were developed according to existing standards for the use of force, it would be deeply problematic. In other words, perfect adherence to existing federal, state and local policies in most jurisdictions–as they currently stand– would be dangerous and downright scary. Similarly, judging the performance of such a system against existing police practices and performance would be setting a terrifying low bar for the performance of such a system.

Even starting from a set of standards such as those sought by #BLM/CampaignZero and the UN

[19] http://www.joincampaignzero.org/solutions/

HRC, would raise some serious concerns. Much like autonomous weapons systems for military applications, it is doubtful that systems would perform at anything like the desired levels, and there are challenges to establishing exactly what those levels should be. Further, there are fundamental moral questions at stake about whether machines should be permitted to use such force at all.

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

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 a 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 in" to the system, and how those rules ought to be interpreted in actual situations.

Roboticists and HRI designers usually 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 force and lethal force against humans, and thus *deliberately causing harms 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 special circumstances, *e.g.* 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 force and lethal force, and what conditions apply to the initiation and escalation of violent force and lethal 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? As a recent Amnesty International report (2015b) 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, 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. In other words, the United States is currently failing to meet its obligations as State party to United Nations Human Rights Conventions (and additional treaties), to ensure the protection of human rights through establishing appropriate laws and polices for the use of force by law enforcement. 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 failiang 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, which 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 an 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 forced is used, when possible.

The first conclusion to draw from this is that building existing United States use of force standards into a future automated robocop ought to be recognized as deeply irresponsible and dangerous. Indeed, as #BlackLivesMatter and CampaignZero have made clear,[20] 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. Failing to do so means that the United States is in violation of its international obligations, and the conventions and treaties to which the US is signatory.

Given that governmental bodies at the federal, state, and local levels are failing to meet international standards, and the federal government is actively failing to meets its obligations under both the treaties that it has signed and customary law, what would it mean to build a robot according to any of these deficient standards? For the roboticist and HRI designer, it would mean complicity in the failure of the United States to meet its international obligations. It would clearly be irresponsible to develop a robotic system that failed to meet the international standards. Building to local standards would be permissible where those standards are more restrictive than the international standards, but not where they are less restrictive. Building a robot to such standards would effectively be aiding and abetting in the violation of the human rights of all those who could be subject to loss of life and violation of bodily sanctity at the hands of those robots.

2. 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,[21] ICRC,[22] and Amnesty

[20] CampaignZero.org

[21] http://www.ohchr.org/EN/ProfessionalInterest/Pages/UseOfForceAndFirearms.aspx

[22] https://www.icrc.org/en/document/use-force-law-enforcement-operations
https://www.icrc.org/eng/assets/files/other/icrc_002_0943.pdf

International,[23] and the legal 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 to 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.

A. International Standards

In a 1990 meeting in Havana, Cuba, the Eighth United Nations Congress on the Prevention of Crime and the Treatment of Offenders adopted the "Basic Principles on the Use of Force and Firearms by Law Enforcement Officials" which embodies the codified standards on international customary law.[24] Similar principles were endorsed by the United Nations General Assembly in 1979, the "Code

---

[23] http://www.amnesty.nl/nieuwsportaal/rapport/use-force-guidelines-implementation-un-basic-principles-use-force-and-firearms

[24] 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.
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.

Special provisions
9. Law enforcement officials shall not use firearms against persons except in self-defence or defence of others against the imminent threat of death or serious injury, to prevent the perpetration of a particularly serious crime involving grave threat to life, to arrest a person presenting such a danger and resisting their authority, or to prevent his or her escape, and only when less extreme means are insufficient to achieve these objectives. In any event, intentional lethal use of firearms

of Conduct for Law Enforcement Officials."[25] Together these represent the international human rights legal standards for the use of force by law enforcement officials.

Taken together, the principles and articles require that the use of force by police officers in law enforcement to meet a number of specific conditions in order to be lawful: 1) it must be *necessary* to prevent an *imminent grave bodily harm or 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.

Given these requirements, how ought 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?

---

may only be made when strictly unavoidable in order to protect life.
10. In the circumstances provided for under principle 9, law enforcement officials shall identify themselves as such and give a clear warning of their intent to use firearms, with sufficient time for the warning to be observed, unless to do so would unduly place the law enforcement officials at risk or would create a risk of death or serious harm to other persons, or would be clearly inappropriate or pointless in the circumstances of the incident.
11. Rules and regulations on the use of firearms by law enforcement officials should include guidelines that:
(a) Specify the circumstances under which law enforcement officials are authorized to carry firearms and prescribe the types of firearms and ammunition permitted;
(b) Ensure that firearms are used only in appropriate circumstances and in a manner likely to decrease the risk of unnecessary harm;
(c) Prohibit the use of those firearms and ammunition that cause unwarranted injury or present an unwarranted risk;
(d) Regulate the control, storage and issuing of firearms, including procedures for ensuring that law enforcement officials are accountable for the firearms and ammunition issued to them;
(e) Provide for warnings to be given, if appropriate, when firearms are to be discharged;
(f) Provide for a system of reporting whenever law enforcement officials use firearms in the performance of their duty.

Policing unlawful assemblies
12. As everyone is allowed to participate in lawful and peaceful assemblies, in accordance with the principles embodied in the Universal Declaration of Human Rights and the International Covenant on Civil and Political Rights, Governments and law enforcement agencies and officials shall recognize that force and firearms may be used only in accordance with principles 13 and 14.
13. In the dispersal of assemblies that are unlawful but non-violent, law enforcement officials shall avoid the use of force or, where that is not practicable, shall restrict such force to the minimum extent necessary.
14. In the dispersal of violent assemblies, law enforcement officials may use firearms only when less dangerous means are not practicable and only to the minimum extent necessary. Law enforcement officials shall not use firearms in such cases, except under the conditions stipulated in principle 9.

Policing persons in custody or detention
15. Law enforcement officials, in their relations with persons in custody or detention, shall not use force, except when strictly necessary for the maintenance of security and order within the institution, or when personal safety is threatened.
16. Law enforcement officials, in their relations with persons in custody or detention, shall not use firearms, except in self-defence or in the defence of others against the immediate threat of death or serious injury, or when strictly necessary to prevent the escape of a person in custody or detention presenting the danger referred to in principle 9.
http://www.ohchr.org/EN/ProfessionalInterest/Pages/LawEnforcementOfficials.aspx

[25] http://www.ohchr.org/EN/ProfessionalInterest/Pages/LawEnforcementOfficials.aspx

B. 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 from CCTV, police dash-cams, and witness cellphones which later went viral on social media.[26] 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 here 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 judgements, 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 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 making a threatening gesture, or how to properly communicate a willingness to cooperate?

There has been much written on the 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, and in which the police officer might be expecting a threat based on the general appearence and manner of an individual. These types of general impressions, which instead 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

[26] 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.

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 beating of Rodney King by Los Angeles Police was subject to expert analysis during the trial, it was deconstructed frame-by-frame to confirm the police report that King posed a 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. 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, like Google DeepMind to recognize such gestures? Should we try empirically determine how the community in which the robot will be used "read" such gestures? Additionally, there could be socially and culturally 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 Afganistan, 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, blind and those utilizing 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.

Many, if not all, technologies make assumptions about the people who may use them. In most cases, they assume that people will fall within the bounds of "normal" in a broad range 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.[27] Presumably these laws would also require law enforcement robots to recognize the special needs of people with

[27]Americans with Disabilities Act , and International Declarations on Disabled

permanent disabilities. It may also require accommodations for individuals who are clearly suffering from temporary episodes, though they are behaving unpredictably and could 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 in to 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 (Chicago). In at least one case, and deaf man was shot for failing to follow verbal police orders after trying to communicate to the officer that he was deaf.

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. Indeed, context is important, but several police shootings have occurred in playgrounds and even the toy-section of a Wal-Mart, where one would hope that the default assumption would be that a gun was a toy.

More problematically, almost any object cold conceivably be used as a weapon, though not all with the same degree of threat. A stick or hammer can be an effective weapon, though it has 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 no 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.

C. Threat Requires Intention

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 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 to 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. 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 or require violent or lethal force to avert. It might be possible to talk someone out of carry 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.

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. As such, they 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 diffuse 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. Many instances of the use of force by police involve threats tot he police officer themself. A robot may be advantageous in dealing with dangerous individuals due to the fact that the need not act out of fear for their own safety, but this carries with it a requirement to use much less force than potentially lethal force, if there is other person around or being imminently threatened.

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." (Cite) 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." The recognition of a threat requires a human-level understanding of the facts and circumstance, 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 fo making these decisions, and setting the certainty parameters? Indeed, in most real-world cases it is the police officer who makes these discretionary judgments,

often with little accountability. 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.[28]

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. But with such insight and understanding of physical systems, would also come greater understanding of how to interfere with them 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, potential victims might be moved out of the way. And similarly, there would be a responsibility to avoid the use of violent and lethal force, within the capabilities of the robotic system. Much of this relates to the question to which we now turn, that of proportionality.

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.[29] 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,[30] can provide insights into social systems and the human experience, and are sometimes effective in encouraging or discouraging certain

[28]DARPA drones, robot videos....

[29]Peter Winch, (1958) *The Idea of a Social Science and its Relation to Philosophy*, London 1958.

[30]Howard S. Becker (1963) *Outsiders: Studies in the Sociology of Deviance.* New York: The Free Press.

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 a 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 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-fulling or self-defeating their predictions.[31] 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 our robocop. By "locking in" a model of human action into the predictive simulator of our 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 right 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 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,

[31] E.g., 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.

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 shifts and develops. 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 explanation after the fact. This is not true for irrevocable deprivations of rights. Most clearly in the use of lethal force–no appeal can bring back the dead. But it 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 can have irrevocable consequences.

3. How Much Violent and Lethal Force is Appropriate and Proportional to a Given Threat?

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 demand explicitly *human* judgements.

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 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 which 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.[32]

[32]It is thus disconcerting that most police officers in the United States are trained to aim shots for the head or

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 and incapacitating a person is generally sufficient to thwart threats not already set in motion, though that does involve the use of force which could be violent, could result in injury, and also deprives and individual of the freedom of movement–and so the threat posed must be weighed against those factors.

There is a technical and moral issue here regarding whether an artificial system can make the type of value judgements that are constitutive of proportionality judgment in the use of force. This problem is even more severe for the use of force in law enforcement, 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 attempt alternatives to violent force, will be quite complex and probabilistic at best.

I have made the similar arguments with regard to proportionality in the use of lethal force by military robots in armed conflict (asaro 2012). 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 infrastucture 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 who violence is directed. Such consideration is not required in armed conflict, but is required in policing.

This is particularly acute problem for trying to design a robocop to deal with individuals suffering from mental illness, as well as individuals with disabilities. Mental illness, and mental episodes, can appear quite confusing an threatening to police officers, and indeed have resulted in a number of police shootings. Because individuals suffering from such episodes may be unable to communicate or response appropriately to police instructions, they may automatically be deemed "uncooperative". Deaf individuals have been shot by police for failing to follow verbal instructions. Similarly, their unpredictable behaviors may easily lead them to be deemed as "threats." Numerous cases involving individuals experiencing episodes have been shot. Indeed, it is often a justification for the use of

---

chest in all cases, or *by default*. This 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. This approach, however, precludes significant proportionality judgments being made once the firearm is drawn. Police in Europe and other countries are trained instead to aim for legs and feet by default.

force for police to claim that they believed the individual was high on drugs–whereas this should probably be viewed as a temporary incapacity rather than a culpable intentional behavior.

Finally, such a system must also be capable of recognizing the de-escalation of a threat. If a suspect throws up their hands and says, "Don't shoot!" or makes similar symbolic acts to tat 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 to proportionality judgment. But it is clear in the law that a human must make such decisions, both because such technologicl solutions are as yet inconceivable, but also because that entails a human who is responsible and accountable for the use of force.

4. 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 its use of force. At the very least this would be transparency with regard to its algorithms and funtioning, as well as logs of its operations and black boxes.

But we cannot really hold robots legally responsible for their actions. Further, it is awkward or impossible to hold programmers responsible. However, this is not unreasonable and probably a good reason for HRI designers and roboticsts to consider a code of ethics that precludes the use of violent and lethal force by robots altogether. Police departments might well be liable to lawsuits due to the use of force by its robots. This might ensure that particular robots are kept up in proper maintenance and software updates. But could they be held liable for civil rights violations if those robots perform in systematically racist or otherwise discriminatory ways?

Individual officer must be accountable for their actions to superiors, but also to the communities which they serve. Community review boards for robots. Analysis of data for the deployment of robots, and logs of interactions with members of the public. Any system flaws in the functioning a law enforcement robot, or systemically unfair deployment ought to be auditable with complaints being investigated and adjudicated where necessary.

Part II: A Bug or a Feature? Embedding Racism in Technologies

It is a commonly held belief that technologies are essentially neutral–that they harbor no biases and are value-neutral. This belief is false, however. The preponderance of research results from the social studies of science and technology demonstrate again and again that technologies are embedded with social values at every level–from low-level design decisions to macro-level social adoption, regulation and implementation of technological infrastructures. These embedded values can exhibit and enforce many forms of bias, including race, class, gender, language and others. In this section I will consider how racial bias in particular might be embedded in automated policing

technologies, at various levels of design and implementation. Such embedded bias could be completely or partially unintentional, or intentional, in the design of the technology.

Ensuring that a technology is truly value-neutral, or free from racial bias requires making this an explicit design goal, and actually testing and evaluating the use of a technology in practice to determine whether that design goal has actually been accomplished. It is not insignificant that establishing such a design goal, and defining how a technology ought to be evaluated in relation to the at goal are themselves highly contentious political issues. Indeed, I would argue that it is precisely because they are political that there need to be a diversity of voices and perspectives involved at all levels of the design, adoption and implementation of technologies.

There are a number of different ways in which racial bias and discrimination could be built into technologies. These range from low-level biases which recognize features of racial difference, and act differently as a result, to higher-level biases that result from analysis of socio-cultural signifiers and context. Examples of such bias could include systems which behave differently in response to certain racialized features, including skin, hair and eye color, as well as hair style, tattoos, etc.; body size and type, as well as age and gender; language, and manners of speech and gesture; clothing and styles of dress; other cultural signifiers such as music, jewelry, text on clothing, associated objects and accessories, cars, bikes, scooters and skateboards, etc. In other words, anything which a system is designed to recognize as a distinguishing feature, or which it learns as such through machine learning techniques. Systems that behave differently in response to these differentiating features could be called discriminatory. This could also include failing to recognize people with various features as people at all, or simply ignoring them.

Depending on what the system is designed to do, recognizing some types of difference might be important to fulfilling its purpose. A robot styling assistant designed to help someone shopping for clothes, or styling their hair, would likely need to recognize various aspects of a persons body, such as shape and build, skin and hair tone, *etc.*, as well as their likely styling interests, judging from their current clothing and hair, and other more complex socio-cultural signifiers. There are of course many different ways for a technology to handle such difference, some of which might be considered socially appropriate, while others would be considered offensive. It is quite challenging to design such systems to behave in socially appropriate ways.

There are already a number of examples of low-level technology designs that embed exclusionary racial bias by failing to work properly for certain groups of people. Such low-level biases include those that rely upon biometric assumptions about potential users that are racially biased or failed to consider how or whether the system would work with some people, e.g. those with dark skin. A good example of this comes from a recent report of the differential performance of the sensors in automated sinks and soap dispensers in bathrooms.[33] These devises use an infrared beam to detect the presence of a hand. They are essentially proximity sensors, which utilize an infrared sensor to pick up reflected IR light when a hand is in close proximity to the emitter and sensor. However, dark

[33] http://mic.com/articles/124899/the-reason-this-racist-soap-dispenser-doesn-t-work-on-black-skin#.84sM1J8X2

skin reflects far less IR light than pale skin. By tuning the sensitivity of the IR detector, and the strength of the IR emitter, the designers of these sensors are making assumptions about the reflectivity of the hands that can operate the faucet. Many such sensors are tuned so as not to be overly sensitive to ambient IR light, coming from other sources than the emitter, and thus require a high degree of reflectivity in the skin of hands which can activate it. Thus, in order to make the device more robust with respect to ambient IR light, the resulting design does not function for people with darker skin complexions.

A very similar problem occurred with the first generation of xbox Kinect gesture camera/controllers.[34] The kinect camera uses an IR camera in conjunction with an RGB camera to create a 3D depth image of the area in front of it. Hand, arm and leg gestures and movements can be recognized by the system. There were, however reports that the system did not work well or properly when used by people with dark skin. Like the faucet sensor, the Kinect camera actively shines an IR light and uses its sensor to detect reflected IR light. Darker surfaces and skin reflect less IR light, and are thus harder to detect. Microsoft claimed there was no such problem, and Consumer Reports trier unsuccessfully to replicate the problem with the Kinect, or with earlier reports of HP'f face recognition software failing to work properly with dark faces.[35]

We might grant that such design choices were completely unintended, and this flaw was unknown to the designers and manufacturers of these faucets. But we could also ask whether the designers of these technologies failed to take a broader enough view of who might use these technologies. Did they test their systems for use by darker hands? Were the potential racial implications of their design decisions were ever considered? Would they have come up if the design teams involved people of color, or if the testing teams and subjects were similarly diverse? Regardless of the intentions and awareness of designers and manufacturers, the resulting technology has a clearly embedded bias with regard to the skin tone of potential hand washers. One hopes that it will be possible to design such sensors to be more racially inclusive, rather than having to design different sensors for different groups, thus essentially recreating segregated washroom facilities and drinking fountains.

Of course, there are also clear examples where technologies are intentionally "tuned" to favor lighter skin over darker skin. This issue has been documented in the case of color film stock.[36] At the introduction of color film in the film industry, there were limitations in the dynamic range of film stock and developing processes to render detail in pale faces relative to dark faces. The industry, being controlled by whites, and seeking to promote white stars, ensured that the new film stocks and processes were tuned to highlight the details of white skin over black skin. As a consequence of these decisions, black faces in color film generally lacked the details and features afforded to white faces. To a large extent, these same dynamic range and contrast issues emerge for analog and digital video. Indeed, some professional digital video cameras include presets that are tuned to difference complexions.

---

[34] http://www.pcworld.com/article/209708/Is_Microsoft_Kinect_Racist.html

[35] http://www.businessinsider.com/microsofts-kinect-has-trouble-recognizing-dark-skinned-faces-2010-11

[36] http://www.cjc-online.ca/index.php/journal/article/view/2196

http://www.buzzfeed.com/syreetamcfadden/teaching-the-camera-to-see-my-skin#.id4VzgqB9

Part IV: Enacting Structural Racism through Technology

While racism is most recognizable in its overt and egregious manifestations, it also exists within persistent and systemic forms that are much more difficult to recognize, challenge and eliminate. In this section of this paper, I will consider how even a robocop that followed use of force guidelines perfectly, and was completely free of any embedded racism of the sort described in the previous section, could still be used to enact and replicate systemic racism.

As mentioned at in the previous sections, there are numerous risks to allowing social statistics and data driven techniques to guide technological design. What might make sense from a narrow engineering perspective may run counter to social norms, values, morality and law. Data-driven policing is a clear example of this problem, where using crime statistics to set law enforcement policies can lead to community-level discrimination. And the growing area of predictive policing takes this to the next level as a broad range direct and indirect traits are could be used to effect racial bias in automated systems, either intentionally or unintentionally.

It is clear from research into data-driven policing policy that using crime statistics to identify "high-crime" areas and subject these to higher levels of policing, and/or more aggressive policing tactics, creates a self-fulfilling prophecy.[37] Given an existing history of racially biased policing, resulting in greater police presence in communities of color, it is easy to use crime statistics to show that there are higher rates of arrests and convictions among people of color. Higher levels of policing result in more stops of people of color, which in turn result in more arrests and convictions. Similarly, more aggressive policing techniques such as "stop-and-frisk" can result in more interactions with people of color, relative to the general population. All of this functions despite data showing that whites are actually more likely to violate laws, e.g. drug possession, and people of color, despite it being much more likely that people of color are arrested and convicted for drug possession.

The same is true for the use of violent and lethal force by police. Because people of color are stopped more frequently than whites, they are disproportionately likely to become involved in confrontations where the police use violent and lethal force against them.

The use of force, like selective surveillance falls under the category of "descretionary policing" (Joh). That is, many of the interactions with the public that are initiated by police are at their descretion–nobody and no rule has required them to engage an individual in an interaction. Of course, responding to a call from the public or intervening in response to an objectively obvious legal violation, officers are often compelled to act. But in a myriad of day to day decisions about who to interact with, when to intervene, where to follow a case, etc., the officer exercises broad discretionary powers.

[37]Data-drive policing

Such discretionary powers are known to be highly susceptible to the psychological bias of individual police officers, both conscious and unconscious. Many times officers are looking for anything "out of the ordinary," or anything that fits their preconceived notions of what is "suspicous." Black people in white neighborhoods are much more likely to be perceived as suspicious, because they deviate from the norm. However, white people in black neighborhoods may not be similarly viewed as suspicious, especially when they are given deference by the conscious or unconcious racial bias of an officer. Thus, a black person might be more likely to be pulled over for driving an expensive car, because that is perceived as atypical and thus suspicious, while a black person driving a deteriorated car might also be deemed suspicious as they are perceived to be more likely to engage in various illegal behaviors. This is how discretionary powers can provide cover for racist policing.

It is tempting, at this point, to wish for a technological solution that would introduce racial equality into these discretionary choices. One might hope that automation technologies would level the playing field and treat individuals more equanamously across racial categories. However, when we look to other types of automation technologies, we find the opposite to be true, and that automated decisions process often amplify and exacerbate existing racia inequalities, rather than eliminate them.

One reason this happens is due to indirect or proxy variables. Consider automated systems for credit rating and lending, where there are clear legal restrictions on using race as factor in determining loan eligibility and rates. While banks cannot use race *directly* in the automated decision processes, they can use a number of other demographic and geographic factors. It has been shown (cite) that for most of the individuals in a given data set, it is possible to correctly identify their race based on a combination of other indicator variables which are not restricted. This set of indicator variables thus act as a proxy for race, allowing automated algorithms to infer race when it is not explicitly indicated, and moreover to effect decisions that impose racial discrimination, even as they can be claimed to not consider or represent racial categories at all. In voting databases, names of felons (known to be disportionately African-American) are used to "clean" voter registrations thus denying voing rights to individuals with similar names, who are also likely to be African-American. Many automated search algorithms also provide racially biased results depending on subtle variations in the names searched, if they coincide with racially distinction spellings (Pasquale, spelling of names). IN mortgage approval software, it is quite easy to implement automated approval and rate-setting algorithms that make racially biased decision based on geographic data. This is because housing policies and social behavior has created racially segregated communities, and thus using an address as factor in evaluating credit-worthiness is, in many cases at least, a good proxy for race. Similarly, much on-line behavior including cites visited, purchases made, and social media networks, can quickly triangulate racial identity and other characteristics, even where these are never explicitly provided.

It is thus necessary to ensure that not only is race not made an explicit factor in automated decision processes, but also that it is not indirectly implemented by proxy. Again, given that this may result as an unintended consequence of implementing an algorithm, it is necessary to deliberately look for and eliminate such bias.

It should not be surprising that the technological issues just discussed map rather closely to many of

the issues of structural racism. Namely, the fact that communities, families and individuals of color are systematically denied access to housing, education, and financing, are due to self-replicating patterns of discrimination and segregation. These are all instances of structural, or infrastructural, racism. There are other examples, such as building the bus underpasses to low for public buses as a way to exclude poor and minority populations from visiting the beach[38] or from moving to certain suburbs in Atlanta.[39]

Yet another example of racial bias inherent in technologies that are assumed to be neutral is illustrated by a recent case in which Google's automatic image annotation system mistakenly labelled African-American faces as "gorillas" in images.[40] Whatever the computational and structural issues that causes this specific case might have been, the racist implications of this error in automated tagging is immediately clear to humans. That is, even if such an error is statistically likely, it has serious social implications that put a greater responsibility on the automated systems to avoid such errors.

While the gorilla-tagging incident did not rely upon incorrectly labelled training examples, there are serious risks of incorporating such data into automated systems. Indeed, the big data techniques employed by Google in their auto-completion algorithm is rife with racism.[41] Because the algorithm collects the most frequently submitted queries, it offers a reflection of statistically popular racist sentiments. For example, by typing "why do black people..." the autocompletion function will suggest finishing your query with "say ax" and "like fried chicken", thus fulfilling stereotypical expectations. This is not limited to racial stereotypes, and typing "why do women..." will produce "cheat", as will "why do men..." All of which goes to show that statistically likely behaviors are not necessarily socially desirable, and we should be careful and conscientious about any systems which automate meaningful decision making based on such data.

This type of data-driven method is likely to be used for a broad range automated decision-making. Which raises a set of issues around notions of social norms and deviance. There are, in fact, numerous ways to embed racism in technologies that are more indirect, less obvious, and much harder to hold designers and manufacturers accountable for, which will be considered in the next section. At this point, I simply wish to reiterate the point that if we want to develop technologies that are not discriminatory in nature, it is essential that we make this an explicit part of the design and evaluation of technologies. It is not enough that the designers and testers do not desire or seek out discriminatory effects from their technologies. We can only expect fairness and equal treatment from technological systems that are deliberately designed to achieve such effects, are evaluated according to those values, and are actively held accountable when they fail or fall short of the established

[38] Langdon Winner, Autonomous Technology, Chapter

[39] http://www.slate.com/articles/news_and_politics/politics/2014/01/atlanta_s_snow_fiasco_the_real_problem_in_the_south_isn_t_weather_it_s_history.html

[40] http://blogs.wsj.com/digits/2015/07/01/google-mistakenly-tags-black-people-as-gorillas-showing-limits-of-algorithms/

[41] http://www.buzzfeed.com/miriamberger/googles-autocomplete-has-some-pretty-racist-thing#.ul3Jg8917 http://www.dailymail.co.uk/sciencetech/article-2326101/Is-Google-making-RACIST-researchers-claim-auto-complete-function-perpetuates-prejudices.html

ideals. This is especially true as standards of social acceptance, inclusivity, and equality rise. That is to say that as the social values we wish to see in our technologies evolve, so too must the technologies. It would be to lock-in certain values, or the standards for their evaluation, in ways that would limit moral and social progress. The flip-side of that flexibility, of course, that regressive values and standards can also be introduced in new technologies.

There are also clear examples where racism is intentionally built into technologies. In many data-driven applications, including credit ratings and loan approvals, are required under law not to be racially discriminatory.[42] As a result, these algorithm cannot explicitly consider race. But while this not be a field in database, it is not difficult to determine race from other variables that are allowed to be used. Those variables thus become proxies for race. An individual's name, as well as what neighborhood they live in, provide strong indicators of race, as does the name and a combination of factors such as schools attended, patterns of travel and purchases, etc.[43]

There has been a growing practice of purging state voter registries in the United States using databases of felons, immigrants and other who are claimed to be ineligible to vote.[44] In many cases, the names in the databases are "permutated" to give variants, e.g. Rich and Dick for Richard. But due to the high ratios of African-American names in felon databases, relative to the population, and hispanic names in immigration databases, this practice clearly disproprtionately affects those communities. Thus it is possible, in the name of limiting voter fraud, to disenfranchise large numbers of people in specific minority communities through such database practices. While it can be claimed that this is not an intentionally racist practice, it is clear the the practice has racially discriminatory effects–it is not a flaw or bug in the system but a feature desired by those ordering and approving such purges. It is also a good example of how seeming neutral technological processes, in this case purging potential ineligible voter from voter registries, can enact systemic racism.

If our robocop is programmed to identify "suspicious" persons or behavior, what exactly would it be looking for? It would seem that there would be a risk of embedding the prejudices of designers into systems that are trying to find such persons. How ought we determine what counts as "suspicious"? Certain manners of dress or cars that "stick out"? Certain types of behavior that are not themselves illegal but that pick out "undesirable or suspicious types," such as loitering or boisterous talking? Will these be rules that engineers come up with from talking to experts such as police? Will these be based in data-driven processes, by analyzing sets of mug shots, or images of people in public that have been tagged on the internet, or tagged by "experts"? What kind of pattern recognition and machine learning techniques might be used, and how might the tagging already reflect racial bias and prejudice? Of course, there is already considerable discretion for police officers to stop and question whomever they deem suspicious, which provides ample room from racial discrimination.[45] Given that the data sets from which machine learning of categories of suspicious persons and

---

[42] Pasquale, Black Box Society Chapter ??

[43] http://heinonline.org/HOL/LandingPage?handle=hein.journals/nylr79&div=33&id=&page=

[44] http://projects.aljazeera.com/2014/double-voters/index.html, http://patch.com/california/lakeelsinore-wildomar/voter-purge-a-racist-republican-effort-or-smart-fraud831f7e5503,

[45] Joh, Discretionary sureveillance.

behavior are likely to be drawn from historical examples, we will now turn to a consideration of that could very easily replicate institutionalized forms of racism.

Part V: Summary and Conclusions

And finally, I conclude with a summary of the most critical issues facing the reform of standards for the use of violent and lethal force by police, the automation of the use of violent and lethal force by machines, and the overarching necessity for reliable systems of accountability at multiple levels.

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 (cite). 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.

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.

Clearly, and 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 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. Building such standards into a HRI system would amount to enabling and perpetuating serious deprivations of human rights under international law. 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 US 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.

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.

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.

One the most significant problems standing in the way of racially just policing is accountability. Insofar as police officers are not accountable to their superiors or the public in terms of transparency and accuracy for the reports of their interactions with members of the public, especially when violent and lethal force is used or death results, there can be no broad based sense of legitimacy or justice in many cases, or trust from members of the public who are discriminatred against with impunity. Accountability is a multi-layer requirement, which includes not only disclosure of incidents, but transparency in the review process, and full criminal liability for officers who violate the law in their use of force.

Like police dash-cams and body-cams, the data trails such systems will generate provide an opportunity for transparency. But that will still be subject to interpretation, and require oversight. A robocop which might also violate the rights of citizens in its use of force presents a more complicated accountability problem. On the one hand we might be able to design low-level racist prejudices out of the system. However, that does not preclude the systeic forms of racism that may result from how those systems get deployed. Still, they should provide the kind of data that would make accountability possible, but only if there are oversight bodies that have access to that data and use it to diminish racial and other forms of discriminiation in the operation and *effects* of deploying

such technologies. It is not reasonable to expect this to happen on its own, or without oversight with the authority to elect what technologies will be deployed, how they will operate, and when and where they will be deployed.

As law enforcement technologies become more sophisticated, the ability of the public to scrutinize their operation and hold it accountable is threatened. As systems become more complex, experts become more empowered to speak about their operation, and non-expert publics are excluded from discussions and decisions.[46] This problem of expertise poses a serious concern for the future development of many types of law enforcement technologies, many of which will face legitimacy crises if they are adopted with little or no community participation or understanding of their functioning.

Technology can be responsive to human needs and values, but only if they are designed to do so, and are continually evaluated and improved in order to do so. Thus, black lives could matter to robocop, but only if we do the hard work of ensuring that it is designed to do so, actively monitor and evaluate law enforcement technologies, and ensure the use and effects of those technologies actually do, in fact, respect the lives of all people.

[46] Asaro 2000 Participatory Design