Updated submission: January 29, 2020

Submission paper:

Hayley Ramsay-Jones, Soka Gakkai International, member of the Campaign to Stop Killer Robots

UN Special Rapporteur on contemporary forms of racism, racial discrimination, xenophobia and related intolerance, workshop on the impact of new information technologies on racial equality

October 17, 2019

Racism and Fully Autonomous Weapons

Introduction

The rise of artificial intelligence is largely due to an increase in power, memory and speed of computers, and the availability of large quantities of data about many aspects of our lives.[i] Through the commercial application of big-data, we are increasingly being sorted into different classifications and stereotypes. In its most benign form, this stereotyping is being used to sell us products via targeted advertising, however, in its most egregious application, we see the weaponization of new information technologies utilize similar classifications based on biased algorithms, to which the consequences for certain communities could be deadly.

In this paper I focus on fully autonomous weapons that are currently being developed for military and law enforcement purposes; and their potential threat to the human rights of marginalized communities, in particular persons of color intersectionally[ii]. This paper will also consider the systemic nature of racism and how racism would be reinforced and perpetuated by fully autonomous weapons.

Racism in Artificial Intelligence

Fully autonomous weapons can select and attack targets without meaningful human control, they operate based on algorithms and data analysis programming. In essence, this means that machines would have the power to make life-and-death decisions over human beings.

The trend towards more autonomy in weaponry without adequate human oversight is alarming especially when we know that digital technologies are not racially neutral. Moreover, when it comes to artificial intelligence (AI) there is an increasing body of evidence that shows that racism operates at every level of the design process and continues to emerge in the production, implementation, distribution and regulation. In this regard AI not only embodies the values and beliefs of the society or individuals that produce them but acts to amplify these biases and the power disparities.[iii]

One example of racism manifesting in AI is the under-representation problem in science, technology, engineering and mathematics (STEM) fields, which in itself is a manifestation of structural racism and patriarchy in western society. Technologies in the west are mostly

developed by white males, and thus perform better for this group. A 2010 study[iv] by researchers at the National Institute of Standards and Technology (NIST) and the University of Texas, found that algorithms designed and tested in East Asia are better at recognizing East Asians, while those designed in Western countries are more accurate at detecting Caucasians. Similarly, sound detecting devices perform better at detecting male, Anglo-American voices and accents, as opposed to female voices, and non-Anglo-American accents.

Research by Joy Buolamwini,[v] reveals that race, skin tone and gender are significant when it comes to facial recognition. Buolamwini demonstrates that facial recognition software recognizes male faces far more accurately than female faces, especially when these faces are white. For darker-skinned people however the error rates were over 19%, and unsurprisingly the systems performed especially badly when presented with the intersection between race and gender, evidenced by a 34.4% error margin when recognizing dark-skinned women.

Despite the concerning error rates in these systems, commercially we already see adaptations of faulty facial recognition systems being rolled out in a variety of ways from soap dispensers to self-driving cars. The issue here is what happens if law enforcement and national security become reliant on a system that can recognize white males with just 1% error rate yet fails to recognize dark-skinned women more than one-third of the time?

These types of applications of new information technology fail people of color intersectionally at a disturbing rate. The fact that these systems are commercially available reveals a blatant disregard for people of color, it also positions "whiteness"[vi] as the norm, the standard for objectivity and reason. These applications of new information technology including their weaponization favors whiteness at the expense of all others, it is not merely a disempowerment but an empowerment. In real terms, racism bolsters white people's life chances. [vii]

As we all grew up in a white-dominated world it is not surprising that the vast majority of white people operate within, benefit from and reproduce a system that they barely notice. This is a long-held reality and it is a fundamental problem that we now see infiltrate technology.

Historical or latent bias in data is another issue, this is created by frequency of occurrence, for example in 2016 an MBA student named Rosalia[viii] discovered that googling "unprofessional hairstyles for work" yielded images of mainly black women with afro-Caribbean hair, conversely when she searched "professional hair" images of mostly coiffed white women emerged, similar google search results are still seen today. This is due to machine learning – algorithms; it collects the most frequently submitted entries and therefore reflects statistically popular racists sentiments. These learnt biases are further strengthened, thus racism continues to be reinforced.

A more perilous example of this is in data-driven, predictive policing that uses crime statistics to identify "high crime" areas and then subjects these areas to higher and often more aggressive levels of policing. Crime happens everywhere, however when an area is over-policed such as communities of color that results in more people of color being arrested and flagged as "persons of interest" thus the cycle continues.

In 2017, Amnesty International launched a report called "trapped in the Matrix",[ix] the report highlighted racially discriminatory practices by the UK police force and their use of a database

called the "Gangs Matrix" which inputs data on "suspected" gang members in London. As of October 2017, there were 3,806 people on the Matrix, 87% of those are from black, Asian and minority ethnic backgrounds and 78% are black, a disproportionate number given that the police's own figures show that only 27% of those responsible for serious youth violence are black.

Amnesty stated that some police officers in the UK have been acting like they are in the "Wild West", making false assumptions about people based on their race, gender, age and socioeconomic status. As a result, individuals on the Matrix database are subject to chronic over-policing. With black people six times more likely to be stopped and searched than white people, and ten times more likely to be convicted of drug-related offenses.

This system not only interferes with their right to privacy, Amnesty claims that the police often share the Matrix with other local agencies such as job centers, housing associations, social services, schools and colleges. In several cases, this has led to devastating impacts on people's social and economic lives because they are listed as "nominal" gang members, a label which is deliberately vague and stigmatizing.

The nature of systemic racism means that it is embedded in all areas of society, the effects of this type of oppression doesn't easily dissipate. Through the continual criminalization and stigmatization of people of color, systemic racism operates by creating winners and losers regardless of what people actually do. This is also the way that it redistributes opportunities and resources based on nothing other than privilege.

Given that the UK, as well as five other countries[x] are developing fully autonomous weapons to target, injure and kill based on data-inputs and pre-programmed algorithms, we can see how long-standing inherent biases, pose an ethical and human rights threat. Where some groups of people will be vastly more vulnerable than others, fully autonomous weapons would not only act to further entrench already existing inequalities but could exacerbate them and lead to deadly consequences.

Legalities

As AI technology advances, the question of who will be held accountable for human rights abuses is becoming increasingly urgent. Machine learning and AI, effect a range of human rights including privacy, freedom of expression, freedom of assembly, the right to non-discrimination and equality, the right to life and the right to human dignity.

Holding those responsible for the unlawful killings of people of color by law enforcement and the military is already a huge challenge in many countries, however, this issue would be further impaired if the unlawful killing was committed by a fully autonomous weapon. Who would be held responsible: the programmer, manufacturer, commanding officer, or the machine itself? Lethal force by these weapons would make it even easier for people of color to be at the mercy of unlawful killings and far more difficult to obtain justice for victims of color and their families.

Conclusion

According to Reni Eddo-Lodge racism perpetuates partly through malice, carelessness and ignorance, it acts to quietly assist some, while hindering others.[xi] It is within this framework that we must grapple with race and the weaponization of new information technologies. In this regard, we should ask ourselves who controls these technologies and what do they think they know about the people they are "classifying"? What are the politics of these relationships and the deeply-rooted systemic forms of discrimination? Who benefits from these technologies and how?

There is a long history of people of color being experimented on for the sake of scientific advances from which they have suffered greatly but do not benefit. An example of this is from James Marion Sims, known as the father of gynecology for reducing maternal death rates in the US, in the 19th century. He conducted his research by performing painful and grotesque experiments on enslaved black women. "All of the early important reproductive health advances were devised by perfecting experiments on black women,".[xii] Today, the maternal death rate for black women in the US is three times higher than it is for white women.

Thus, when it comes to new information technology, facial recognition systems, algorithms and automated and interactive machine decision-making, communities of color are often both deprived of their benefits and subjected to their consequences. This paradox where science is inflicted on communities of color rather than aided by it must be addressed.

We must be vigilant against deeply rooted social problems taking root in the technical infrastructure that we create. We must work towards a zero policy on racism in technology, and not weaponize racism in technology. If racism and killer robots are allowed to co-exists these weapons will be used discriminately against people of color and other marginalized groups.

For these and many other ethical, moral, human rights, legal and humanitarian reasons the Campaign to Stop Killer Robots, numerous governments, regional groups, tech workers, experts, scholars and the UN Secretary-General are all calling for a legally binding instrument to prohibit fully autonomous weapons [xiii]

We call on the Special Rapporteur on contemporary forms of racism, racial discrimination, xenophobia and related intolerance to condemn fully autonomous weapons and the human rights threat they pose to people of color; and to support a prohibition treaty that will preserve meaningful human control over the use of force and prohibit fully autonomous weapons.[xiv]

---

[i] Noel Sharkey, International Committee for Robot Arms Control

[ii] For the purposes of this paper I have focused on race, skin tone and gender. Fully autonomous weapons also pose a threat based on people's religion, ethnicity, ability, sexual orientation, gender expression and class among others.

[iii] Peter Asaro,"Will #BlackLivesMatter to Robocop?"

[iv] P. Phillips, Hyeonjoon Moon,"An other-race effect for face recognition algorithms"

[v] Joy Buolamwini, Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification

[vi] http://www.aclrc.com/whiteness

[vii] Reni Eddo-Lodge, Why I no longer talk to white people about race

[viii] https://twitter.com/HhereroRocher/status/717457819864272896]

[ix] Amnesty International, Trapped in the Matrix

[x] The USA, China, Israel, South Korea, Russia and the UK, The Campaign to Stop Killer Robots

[xi] Reni Eddo-Lodge, http://renieddolodge.co.uk/

[xii] Harriet A. Washington, "Medical Apartheid"

[xiii] The Campaign to Stop Killer Robots

[xiv] The Campaign to Stop Killer Robots, Key Elements of a Treaty on Fully Autonomous Weapons (Annexed)

Hayley Ramsay-Jones
Soka Gakkai International member of the Campaign to Stop Killer Robots

Updated submission: January 29, 2020

Racism and Fully Autonomous Weapons

Introduction

The rise of artificial intelligence is largely due to an increase in power, memory and speed of computers, and the availability of large quantities of data about many aspects of our lives.[i] Through the commercial application of big-data, we are increasingly being sorted into different classifications and stereotypes. In its most benign form, this stereotyping is being used to sell us products via targeted advertising, however, in its most egregious application, we see the weaponization of new information technologies utilize similar classifications based on biased algorithms, to which the consequences for certain communities could be deadly.

In this paper I focus on fully autonomous weapons that are currently being developed for military and law enforcement purposes; and their potential threat to the human rights of marginalized communities, in particular persons of color intersectionally[ii]. This paper will also consider the systemic nature of racism and how racism would be reinforced and perpetuated by fully autonomous weapons.

Racism in Artificial Intelligence

Fully autonomous weapons can select and attack targets without meaningful human control, they operate based on algorithms and data analysis programming. In essence, this means that machines would have the power to make life-and-death decisions over human beings.

The trend towards more autonomy in weaponry without adequate human oversight is alarming especially when we know that digital technologies are not racially neutral. Moreover, when it comes to artificial intelligence (AI) there is an increasing body of evidence that shows that racism operates at every level of the design process and continues to emerge in the production, implementation, distribution and regulation. In this regard AI not only embodies the values and beliefs of the society or individuals that produce them but acts to amplify these biases and the power disparities.[iii]

One example of racism manifesting in AI is the under-representation problem in science, technology, engineering and mathematics (STEM) fields, which in itself is a manifestation of structural racism and patriarchy in western society. Technologies in the west are mostly developed by white males, and thus perform better for this group. A 2010 study[iv] by researchers at the National Institute of Standards and Technology (NIST) and the University of Texas, found that algorithms designed and tested in East Asia are better at recognizing East Asians, while those designed in Western countries are more accurate at detecting Caucasians. Similarly, sound detecting devices perform better at detecting male, Anglo-American voices and accents, as opposed to female voices, and non-Anglo-American accents.

Research by Joy Buolamwini,[v] reveals that race, skin tone and gender are significant when it comes to facial recognition. Buolamwini demonstrates that facial recognition software recognizes male faces far more accurately than female faces, especially when these faces are white. For darker-skinned people

however the error rates were over 19%, and unsurprisingly the systems performed especially badly when presented with the intersection between race and gender, evidenced by a 34.4% error margin when recognizing dark-skinned women.

Despite the concerning error rates in these systems, commercially we already see adaptations of faulty facial recognition systems being rolled out in a variety of ways from soap dispensers to self-driving cars. The issue here is what happens if law enforcement and national security become reliant on a system that can recognize white males with just 1% error rate yet fails to recognize dark-skinned women more than one-third of the time?

These types of applications of new information technology fail people of color intersectionally at a disturbing rate. The fact that these systems are commercially available reveals a blatant disregard for people of color, it also positions "whiteness"[vi] as the norm, the standard for objectivity and reason. These applications of new information technology including their weaponization favors whiteness at the expense of all others, it is not merely a disempowerment but an empowerment. In real terms, racism bolsters white people's life chances. [vii]

As we all grew up in a white-dominated world it is not surprising that the vast majority of white people operate within, benefit from and reproduce a system that they barely notice. This is a long-held reality and it is a fundamental problem that we now see infiltrate technology.

Historical or latent bias in data is another issue, this is created by frequency of occurrence, for example in 2016 an MBA student named Rosalia[viii] discovered that googling "unprofessional hairstyles for work" yielded images of mainly black women with afro-Caribbean hair, conversely when she searched "professional hair" images of mostly coiffed white women emerged, similar google search results are still seen today. This is due to machine learning – algorithms; it collects the most frequently submitted entries and therefore reflects statistically popular racists sentiments. These learnt biases are further strengthened, thus racism continues to be reinforced.

A more perilous example of this is in data-driven, predictive policing that uses crime statistics to identify "high crime" areas and then subjects these areas to higher and often more aggressive levels of policing. Crime happens everywhere, however when an area is over-policed such as communities of color that results in more people of color being arrested and flagged as "persons of interest" thus the cycle continues.

In 2017, Amnesty International launched a report called "trapped in the Matrix",[ix] the report highlighted racially discriminatory practices by the UK police force and their use of a database called the "Gangs Matrix" which inputs data on "suspected" gang members in London. As of October 2017, there were 3,806 people on the Matrix, 87% of those are from black, Asian and minority ethnic backgrounds and 78% are black, a disproportionate number given that the police's own figures show that only 27% of those responsible for serious youth violence are black.

Amnesty stated that some police officers in the UK have been acting like they are in the "Wild West", making false assumptions about people based on their race, gender, age and socioeconomic status. As a result, individuals on the Matrix database are subject to chronic over-policing. With black people six times more likely to be stopped and searched than white people, and ten times more likely to be convicted of drug-related offenses.

This system not only interferes with their right to privacy, Amnesty claims that the police often share the Matrix with other local agencies such as job centers, housing associations, social services, schools and colleges. In several cases, this has led to devastating impacts on people's social and economic lives because they are listed as "nominal" gang members, a label which is deliberately vague and stigmatizing.

The nature of systemic racism means that it is embedded in all areas of society, the effects of this type of oppression doesn't easily dissipate. Through the continual criminalization and stigmatization of people of color, systemic racism operates by creating winners and losers regardless of what people actually do. This is also the way that it redistributes opportunities and resources based on nothing other than privilege.

Given that the UK, as well as five other countries[x] are developing fully autonomous weapons to target, injure and kill based on data-inputs and pre-programmed algorithms, we can see how long-standing inherent biases, pose an ethical and human rights threat. Where some groups of people will be vastly more vulnerable than others, fully autonomous weapons would not only act to further entrench already existing inequalities but could exacerbate them and lead to deadly consequences.

Legalities

As AI technology advances, the question of who will be held accountable for human rights abuses is becoming increasingly urgent. Machine learning and AI, effect a range of human rights including privacy, freedom of expression, freedom of assembly, the right to non-discrimination and equality, the right to life and the right to human dignity.

Holding those responsible for the unlawful killings of people of color by law enforcement and the military is already a huge challenge in many countries, however, this issue would be further impaired if the unlawful killing was committed by a fully autonomous weapon. Who would be held responsible: the programmer, manufacturer, commanding officer, or the machine itself? Lethal force by these weapons would make it even easier for people of color to be at the mercy of unlawful killings and far more difficult to obtain justice for victims of color and their families.

Conclusion

According to Reni Eddo-Lodge racism perpetuates partly through malice, carelessness and ignorance, it acts to quietly assist some, while hindering others.[xi] It is within this framework that we must grapple with race and the weaponization of new information technologies. In this regard, we should ask ourselves who controls these technologies and what do they think they know about the people they are "classifying"? What are the politics of these relationships and the deeply-rooted systemic forms of discrimination? Who benefits from these technologies and how?

There is a long history of people of color being experimented on for the sake of scientific advances from which they have suffered greatly but do not benefit. An example of this is from James Marion Sims, known as the father of gynecology for reducing maternal death rates in the US, in the 19th century. He conducted his research by performing painful and grotesque experiments on enslaved black women. "All of the early important reproductive health advances were devised by perfecting experiments on black women,".[xii] Today, the maternal death rate for black women in the US is three times higher than it is for white women.

Thus, when it comes to new information technology, facial recognition systems, algorithms and automated and interactive machine decision-making, communities of color are often both deprived of their benefits and subjected to their consequences. This paradox where science is inflicted on communities of color rather than aided by it must be addressed.

We must be vigilant against deeply rooted social problems taking root in the technical infrastructure that we create. We must work towards a zero policy on racism in technology, and not weaponize racism in technology. If racism and killer robots are allowed to co-exists these weapons will be used discriminately against people of color and other marginalized groups.

For these and many other ethical, moral, human rights, legal and humanitarian reasons the Campaign to Stop Killer Robots, numerous governments, regional groups, tech workers, experts, scholars and the UN Secretary-General are all calling for a legally binding instrument to prohibit fully autonomous weapons. [xiii]

We call on the Special Rapporteur on contemporary forms of racism, racial discrimination, xenophobia and related intolerance to condemn fully autonomous weapons and the human rights threat they pose to people of color; and to support a prohibition treaty that will preserve meaningful human control over the use of force and prohibit fully autonomous weapons.[xiv]

---

[i] Noel Sharkey, International Committee for Robot Arms Control

[ii] For the purposes of this paper I have focused on race, skin tone and gender. Fully autonomous weapons also pose a threat based on people's religion, ethnicity, ability, sexual orientation, gender expression and class among others.

[iii] Peter Asaro,"Will #BlackLivesMatter to Robocop?"

[iv] P. Phillips, Hyeonjoon Moon,"An other-race effect for face recognition algorithms"

[v] Joy Buolamwini, Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification

[vi] http://www.aclrc.com/whiteness

[vii] Reni Eddo-Lodge, Why I no longer talk to white people about race

[viii] https://twitter.com/HereroRocher/status/717457819864272896]

[ix] Amnesty International, Trapped in the Matrix

[x] The USA, China, Israel, South Korea, Russia and the UK, The Campaign to Stop Killer Robots

[xi] Reni Eddo-Lodge, http://renieddolodge.co.uk/

[xii] Harriet A. Washington, "Medical Apartheid"

[xiii] The Campaign to Stop Killer Robots

[xiv] The Campaign to Stop Killer Robots, Key Elements of a Treaty on Fully Autonomous Weapons (Annexed)

1. Specific asks for 2019 Convention on Conventional Weapons

In their statements, working papers, and related documents, states should aim to:

- **Express concern** regarding the many **legal, ethical and security issues** related to lethal autonomous weapons systems and **revive consideration** of key concerns such as **ethics and morality**, humanitarian impacts, and human rights; and
- **Make explicit** how the necessary level and form of **human control** over weapons systems can be enacted and ensured;
- Express deep concern in the slow pace in addressing the issue at the CCW;
- **Adopt a new CCW mandate to negotiate a legally binding instrument** to prohibit lethal autonomous weapons systems and ensure meaningful human control over the use of force. This mandate should include a significant number of working days each year for 2020 and 2021. If this is not possible within the CCW, other diplomatic options should be explored.

2. What is the Campaign to Stop Killer Robots calling for?

The Campaign to Stop Killer Robots is a **global coalition** of 130 non-governmental organizations from 60 countries working to retain meaningful human control over the use of force by banning the development, production, and use of fully autonomous weapons, known in the Convention on Conventional Weapons (CCW) as lethal autonomous weapons systems. **The campaign calls on all states to:**

- Launch negotiations for a legally-binding instrument that ensures meaningful human control over the use of force and prohibits lethal autonomous weapons systems. In particular, seek a revised mandate at the CCW's annual meeting on 15 November to begin negotiating a new protocol in 2020;
- Commit not to develop or acquire fully autonomous weapons and establish national policies and laws towards this objective, in consultation with civil society and other national stakeholders;
- Specify the necessary human control required over the use of force, and in particular the critical functions of identifying, selecting, and engaging targets.

3. Who else is calling for a prohibition on fully autonomous weapons?

Currently, **30 countries**[1] are explicitly calling for a prohibition on fully autonomous weapons, while the Non-Aligned Movement (NAM) has also called for a prohibition and restrictions on the development and use of fully autonomous weapons. During the August 2018 GGE, Austria, Brazil and Chile proposed a new CCW Mandate "to negotiate a legally-binding instrument to ensure meaningful human control over the critical functions" of weapons systems.

In September 2018 the European Parliament called for the start of negotiations on a ban on lethal autonomous weapon systems. In addition, **German** Foreign Minister Heiko Maas has stated on multiple occasions that Germany wants to ban lethal autonomous weapons systems. On 23 March 2019, **Belgian** Foreign and Defence Minister Didier Reynders also for the first time stated that Belgium aims to prohibit "autonomous weapons capable of killing without any human intervention." In July, the parliamentary assembly of **the Organization for Security and Co-operation in Europe** (OSCE) adopted a declaration that includes a line urging the participating states "to support international negotiations to ban lethal autonomous weapons."

There is also clear public concern. In an **IPSOS survey** released in January 2019, more than three in every five people (61%) in 26 countries stated their opposition to the development of weapons systems that would select and attack targets without human intervention. Two-thirds (66%) of those opposed to lethal autonomous weapons systems were most concerned that they would "cross a moral line because machines should not be allowed to kill."

Additionally, over 4500 **Artificial Intelligence experts**, and 116 CEO's from robotics companies have warned against these weapons and called on the United Nations to take action. More than 240 **tech companies** and over 3200 individuals pledged to never develop, produce or use of lethal autonomous weapon systems.

Finally, the **International Committee of the Red Cross** has called on states to establish internationally agreed limits on autonomy in weapon systems, which address legal, ethical and humanitarian concerns. **UN Secretary-General** Guterres has also called lethal autonomous weapons "morally repugnant and politically unacceptable", and has urged states to negotiate a ban on these weapons.

[1] Algeria, Argentina, Austria, Bolivia, Brazil, Chile, China*(use only), Colombia, Costa Rica, Cuba, Djibouti, Ecuador, Egypt, El Salvador, Ghana, Guatemala, Holy See, Iraq, Jordan, Mexico, Morocco, Namibia, Nicaragua, Pakistan, Panama, Peru, State of Palestine, Uganda, Venezuela, and Zimbabwe.

KEY ELEMENTS OF A TREATY

ON FULLY AUTONOMOUS WEAPONS

CAMPAIGN TO STOP
KILLER ROBOTS

KEY ELEMENTS OF A TREATY ON FULLY AUTONOMOUS WEAPONS

The increasing technological capacity for autonomy in weapons systems raises a host of moral, legal, accountability, technological, and security concerns. Weapons systems that select and engage targets without meaningful human control—known as fully autonomous weapons, lethal autonomous weapons systems, or killer robots—would cross the threshold of acceptability and should be prevented and prohibited through new international law.

The Campaign to Stop Killer Robots is calling for a legally binding instrument to address such emerging technology by preserving meaningful human control over the use of force. The instrument should apply to the range of weapons systems that select and engage targets on the basis of sensor inputs, that is, systems where the object to be attacked is determined by sensor processing, not by humans.[1] This broad scope is designed to ensure problematic technology does not escape regulation.

The treaty's restrictions, however, would focus on those systems that contravene the requirement of meaningful human control. It would use a combination of prohibitions and positive obligations effectively to ban systems that amount to, or are used as fully autonomous weapons. While specific language and content would have to be worked out during multilateral discussions and treaty negotiations, the final instrument should incorporate the key elements identified in this paper.

[1] For more on this categorization, see Richard Moyes, Article 36, "Target Profiles," August 2019, http://www.article36.org/wp-content/uploads/2019/08/Target-profiles.pdf, p. 3.

This paper examines the concept of meaningful human control, which would be central to the new treaty or protocol. It then proposes three types of core obligations:

- A general obligation to maintain meaningful human control over the use of force;

- Prohibitions (i.e., negative obligations) on weapons systems that select and engage targets and by their nature pose fundamental moral or legal problems; and

- Specific positive obligations to help ensure that meaningful human control is maintained in the use of all other systems that select and engage targets.

THE CONCEPT OF MEANINGFUL HUMAN CONTROL

The proposed legally binding instrument should focus on meaningful human control because many of the concerns raised by fully autonomous weapons are attributable to the absence of such control. This absence would undermine human dignity to delegate life-and-death determinations to inanimate machines that reduce humans to datapoints yet could not comprehend the value of human life. Such weapons systems would also lack the capacity for human judgment necessary, for example, to weigh the proportionality of an attack, as required under international law. Furthermore, it would be legally difficult and arguably unjust to hold a human liable for the actions of a system operating beyond his or her control.[2]

For these and other reasons, states as well as international and non-governmental organizations have expressed widespread agreement about the need for some form of human control over the use of force. Their choice of terminology and specific views of the human role may differ, but they have identified many of the same factors. Drawing on international discussions and numerous publications, this paper distills the concept of meaningful human control into decision-making, technological, and operational components.[3]

[2] For more information on the problems of fully autonomous weapons, see Human Rights Watch and the Harvard Law School International Human Rights Clinic, *Making the Case: The Dangers of Killer Robots and the Need for a Preemptive Ban* (2016), https://www.hrw.org/report/2016/12/09/making-case/dangers-killer-robots-and-need-preemptive-ban.

[3] While there are different ways to frame this concept, the phrase “meaningful human control” has many advantages. “Control” is a term widely used in international law and is stronger and broader than the alternatives proposed by a few states, such as intervention and judgment. The qualifier “meaningful” works to ensure that control is substantive rather than superficial and is less context specific or outcome driven than alternatives like appropriate and effective.

DECISION-MAKING COMPONENTS

The decision-making components of meaningful human control give humans the information and ability to make decisions about whether the use of force complies with legal rules and ethical principles. In particular, the human operator of a weapon system should have: an understanding of the operational environment; an understanding of how the system functions, including what it might identify as a target; and sufficient time for deliberation.

TECHNOLOGICAL COMPONENTS

Technological components are embedded features of a weapon system that can enhance meaningful human control. They include: predictability and reliability;[4] the ability of the system to relay relevant information to the human operator; and the ability for a human to intervene after the activation of the system.

OPERATIONAL COMPONENTS

Operational components make human control more meaningful by limiting when and where a weapon system can operate and what it can target. Factors that could be constrained include: the time between a human's legal assessment and the system's application of force; the duration of the system's operation; the nature and size of the geographic area of operation; and the permissible types of targets (e.g., personnel or material).

While none of these components are independently sufficient to amount to meaningful human control, all have the potential to enhance control in some way. In addition, the components often work in tandem. Further analysis of existing and emerging technology could help determine which these or other components should be codified in a legal instrument as prerequisites for meaningful human control.

[4] In general, predictability refers to the degree to which a weapon system operates as humans expect, and reliability refers to the degree to which the system will perform consistently. International Committee of the Red Cross statement under Agenda Item 5(b), Convention on Conventional Weapons Group of Governmental Experts on Lethal Autonomous Weapons Systems, Geneva, March 2019.

CORE OBLIGATIONS OF THE TREATY

The heart of the legally binding instrument should consist of three core obligations: a general obligation along with prohibitions and positive obligations to implement it.

A GENERAL OBLIGATION TO MAINTAIN MEANINGFUL HUMAN CONTROL OVER THE USE OF FORCE

This overarching provision would facilitate compliance with applicable legal and ethical norms by obliging states parties to maintain meaningful human control over the use of force. The generality of the obligation would help avoid loopholes, and the principle it embodies could inform interpretation of the treaty's other provisions. As noted above, most states have already expressed support for a requirement of human control.

The general obligation should focus on control over conduct ("use of force") rather than specific technology. This approach would help future-proof the treaty by obviating the need to predict how technology will develop. The term "use of force" also makes the general obligation applicable to situations of armed conflict and law enforcement.[5]

[5] While the term "use of force" frequently appears in discussions and documents of international humanitarian law and international human rights law, the two bodies of law govern it is somewhat different ways. The new treaty may need to take such differences into account.

PROHIBITIONS ON SPECIFIC WEAPONS SYSTEMS THAT SELECT AND ENGAGE TARGETS AND BY THEIR NATURE POSE FUNDAMENTAL MORAL OR LEGAL PROBLEMS

The treaty should prohibit the development, production, and use of weapons systems that select and engage targets and are inherently unacceptable for ethical or legal reasons. The clarity of the prohibitions would facilitate monitoring and enforcement, and their absoluteness would create a strong stigma against the banned systems.

The new instrument should prohibit weapons systems that by their nature select and engage targets without meaningful human control. The prohibition should cover, for example, systems that become too complex for human users to understand and thus produce unpredictable and inexplicable effects. These complex systems might apply force based on prior machine learning or allow critical system parameters to change without human authorization. Such weapons systems would run afoul of the new instrument's general obligation discussed above.

The prohibitions could also extend to specific other weapons systems that select and engage targets and are by their nature, rather than their manner of use, problematic. In particular, the treaty could prohibit weapons systems that select and engage humans as targets, regardless of whether they operate under meaningful human control.[6] Such systems would rely on certain types of data, such as weight, heat, or sound, to represent people or categories of people. In killing or injuring people based on such data, these systems would contravene the principle of human dignity and dehumanize violence. A prohibition on this category of systems would also encompass systems that, deliberately or unintentionally, target groups of people based on discriminatory indicators related to age, gender, or other social identities.

[6] For more information on such systems and the proposal to prohibit them, see generally Moyes, "Target Profiles."

SPECIFIC POSITIVE OBLIGATIONS TO ENSURE THAT MEANINGFUL HUMAN CONTROL IS MAINTAINED IN THE USE OF ALL OTHER SYSTEMS THAT SELECT AND ENGAGE TARGETS

The new instrument's positive obligations should cover weapons systems that are not inherently unacceptable but that might still have the potential to select and engage targets without meaningful human control. The obligations would require states parties to ensure that weapons systems that select and engage targets are used only with meaningful human control.

The content of the positive obligations should draw on the components of meaningful human control discussed above. For example, the treaty could require that operators understand how a weapon system functions before activating it. It could set minimum standards for predictability and reliability. In addition, or alternatively, the treaty could limit permissible systems to those operating within certain temporal or geographic parameters. In so doing, the positive obligations would help preserve meaningful human control over the use of force and establish requirements that in effect render the use of system operating as fully autonomous weapons unlawful.

OTHER ELEMENTS

While the key elements outlined above are critical to achieving the objectives of the new instrument, other elements should complement them. For example, a preamble should articulate the purpose of the treaty and place it in the context of relevant law. Reporting requirements would promote transparency and facilitate independent monitoring. Detailed verification measures or cooperative compliance mechanisms would help prevent violations of the treaty. Regular meetings of states parties would provide an opportunity to review the status and operation of the treaty, identify implementation gaps, and set goals for the future. Other important elements would include a requirement to adopt national implementation measures and a threshold for entry into force.

This Campaign to Stop Killer Robots briefing paper was prepared by Bonnie Docherty of Human Rights Watch and the Harvard Law School International Human Rights Clinic, with the support of her law students in the Clinic.

Retain meaningful human control over the use of force.
Prohibit fully autonomous weapons.
WWW.STOPKILLERROBOTS.ORG

*****PRELIMINARY DRAFT***3-28-16***DO NOT CITE WITHOUT PERMISSION******

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

Abstract

Introduction

#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

---

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

[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

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 interactions with the public. 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

---

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/)

[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/).

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 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

---

[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.

[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/

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 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

---

[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

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 is 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 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 embedded 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 egregiously violent 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 reproduce systemic racism.

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.

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.

Part I: Automating Police Use of Violent and Lethal Force

Among the various activities the police typically perform, the most morally and politically significant

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

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.

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.

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 the given rules for 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, how those rules ought to be interpreted in actual situations, and whether a machine is capable of meeting the legal and moral requirements for the use of violent and lethal force.

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

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 for their robot? As a recent Amnesty International report[20] 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, while there is no specific federal legislation in place, Supreme Court decisions have set constitutional law standards for the use of violent and lethal force, and the Department of Justice has issued its own guidelines.[21] Most state and local laws and 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

[20] http://www.amnestyusa.org/research/reports/deadly-force-police-use-of-lethal-force-in-the-united-states

[21] http://www.nij.gov/topics/law-enforcement/officer-safety/use-of-force/pages/welcome.aspx

local jurisdiction? Or choose one, or both, of the federal standards? Or allow their customers, local police departments, to choose from the sets of constraints they wish to adhere to? This is a design choice fraught with peril.

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 party to United Nations Universal Declaration of Human Rights,[22] and additional instruments that establish policing standards to ensure the protection of human rights through establishing appropriate laws and polices for the use of force by law enforcement.[23] These failures are as complete and far-reaching as they are distressing. Some U.S. states have failed 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, or even the federal standards which themselves fail to meet the minimal international standards. International law calls for all nations to "establish laws" to ensure the protection of human rights and restrain police in the use of force, while the U.S. federal government has failed to establish any laws on this matter.

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 imminent threat to others, or even pose no significant risk to cause harm in the future, to permitting lethal force as a first resort rather than last, to failing to establish policies and procedures for documenting the use of force and discharge of firearms, to failing to establish inquiries into police actions resulting in death and serious injury, to failing to provide oversight mechanisms for monitoring and reviewing police use of force, or mechanisms to ensure proper and effective training of police in proper standards and procedures. All of these are failures to meet the international guidelines, which only permit the use of force when there is an imminent 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. Beyond that, there are requirements for reporting and reviewing any instances where police use force against citizens.

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,[24] 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

---

[22] http://www.un.org/en/universal-declaration-human-rights/

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

[24] CampaignZero.org

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 international standards. Building to local standards would be permissible where those standards are more restrictive than the minimal 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,[25] ICRC,[26] and Amnesty International,[27] 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.[28] Similar principles were endorsed by the United Nations General Assembly in 1979, the "Code

---

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

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

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

[28] 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 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

of Conduct for Law Enforcement Officials."[29] 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 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?

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.[30] 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 the designer of a robocop, what types of gestures, actions and behaviors should count as "threats," or "willingness to cooperate," and how can they be recognized?

---

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

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

[30] 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.

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 applied by prosecutors in most cases, as well as the fact that prosecutors often work together with police and find it difficult to bring charges in most cases.

It is worth asking 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 interpret the ways in which the public actually interacts with police officers for programming a robocop's reactions, 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, 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 could 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 simplest of gestures, or even compliance 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 eight police officers who were beating and tasering him as he lay face down on the ground,[31] 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 imbuing them with intention, rather than seeing them as normal reactions to being violently struck. It was also crucial in that case that officers held a contextualizing assumption that Mr. King was high on powerful drugs and possessed an almost super-human strength and tolerance for pain, which was not true. 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. Mr. King claimed he did try to evade

[31] https://en.wikipedia.org/wiki/Rodney_King#Beating_with_batons:_the_Holliday_video
Goodwin, Charles. "Professional vision." *American Anthropologist* 96.3 (1994): 606-633.

police in order to avoid a DUI charge that would jeopardize his parole.

We should hope that any police robot would do better than the LAPD with regard to the use of force in such cases. 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 to empirically determine how the community in which the robot will be used "reads" 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 Punjab, Pashtun shepherds carrying rifles in Afganistan, or suburbanites exercising their open-carry rights in Texas.[32]

Beyond recognizing what gestures ought to be considered threatening or compliant in ideal or "normal" circumstances, 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 variety 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, a failure to accomodate such citizens 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 by law in the United States, and now internationally.[33] Presumably these laws would also require law enforcement robots to recognize the special needs of people with permanent disabilities. It may also require accommodations for individuals who are clearly suffering from temporary episodes, though they may be behaving unpredictably and could pose a danger to themselves or those around them.

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

[32] "What to Do" Open Carry PSA, City of Round Rock, Texas, 2015.

[33] Americans with Disabilities Act, and the UN Declaration on the Rights of Disabled Persons

have involved individuals with known mental health issues being shot even when the police were informed of their mental conditions, or even when they were called to give assistance.[34] 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.[35]

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[36] and even the toy-section of a Wal-Mart,[37] where one would hope that the default assumption would be that a gun was a toy. Then there are guns disguised as banal objects, which present another problem.[38]

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.[39] 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

---

[34] http://www.chicagotribune.com/ct-chicago-police-shooting-20151226-story.html, http://www.latimes.com/local/lanow/la-me-ln-lapd-use-of-force-report-20160301-story.html

[35] http://www.huffingtonpost.com/2014/09/23/edward-miller-deaf-man-fatally-shot_n_5868538.html, http://www.usatoday.com/story/news/nation/2014/08/26/krupinski-detroit-police-shooting/14634913/ http://www.theguardian.com/uk/2012/oct/17/police-taser-blind-man-stick

[36] http://www.cnn.com/videos/justice/2015/12/28/tamir-rice-shooting-grand-jury-saw-enhanced-video-casarez-sot-nr.cnn

[37] http://www.cnn.com/2014/09/22/us/ohio-walmart-death/index.html http://www.cnn.com/2014/12/16/justice/walmart-shooting-john-crawford/

[38] http://bgr.com/2016/03/25/smartphone-gun-ideal-conceal/

[39] http://www.thestranger.com/blogs/slog/2015/04/24/22109915/william-wingate-sues-officer-cynthia-whitlatch-and-the-seattle-police-department-alleging-racial-discrimination http://www.cbsnews.com/news/video-shows-south-carolina-deputy-crying-after-shooting-70-year-old-man/

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 Motion interfaces[40]). 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,[41] 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 to the 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 no other person around who is being imminently threatened.

Much of the interpretation of verbal and gestural intentions seems open to differing subjective

---

[40] https://community.leapmotion.com/t/sensor-seems-to-have-trouble-with-darker-skin-color/2351 https://dgoins.wordpress.com/2015/03/21/my-kinect-told-me-i-have-dark-olive-green-skin/

[41] https://www.youtube.com/watch?v=SGxKhUuZ0Rc

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."[42] 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 should a robot wait to act only when there is certainty, or a sufficient degree of certainty? Should HRI designers be the ones responsible for 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 in anticipating threats, but numerous examples of when they get it wrong.

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.

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

[42] http://www.amnestyusa.org/research/reports/deadly-force-police-use-of-lethal-force-in-the-united-states

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.

This potential to model physical systems does not translate a similar potential to predict 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.[43] 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,[44] can provide insights into social systems and the human experience, and are sometimes effective in encouraging or discouraging certain behaviors, or influencing individuals through communication and coercion. But such scientific understanding is not, strictly speaking, predictive of individual behaviors in individual situations. While positivist social scientists have long sought to emulate the precision and predictive powers of the physical sciences, there are fundamental hurdles to doing so. One can argue that this is due to lack of experimental control, imprecise measurement, insufficient conceptual clarity or theoretical understanding, or simply 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.[45] 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. The potential for an interaction to become violent is not itself a reason to initiate or escalate that violence.

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

---

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

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

[45] 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.

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, 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. By drawing upon statistical data models, or narrow sensor data, automated systems use these as proxies for actions, and pass judgement on the proxies rather than the actions and intentions of people. 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?

Deciding how much force is appropriate in the given circumstances, and when and how to escalate the use of force, is known as *proportionality* in the use of force. Again, there are questions of which legal standards to conform to, but also much more challenging technical issues involving how to meet those requirements given that they demand explicitly *human* 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.[46]

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 aiming 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 an 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 the proportionality judgment in the use of force, which has been discussed in the context of autonomous weapons in armed conflict.[47] 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 in many cases. 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!"

---

[46] It is thus disconcerting that most police officers in the United States are trained to aim shots for the head or chest in all cases, or *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.

[47] Asaro, P. (2012). "On Banning Autonomous Lethal Systems: Human Rights, Automation and the Dehumanizing of Lethal Decision-making," Special Issue on New Technologies and Warfare, International Review of the Red Cross, 94 (886), Summer 2012, pp. 687-709.

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.[48] 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 under the international guidelines for police use of force.

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 that effect, the robot must also de-escalate its use of force. Of course, such a robot might get fooled, but it has to provide that opportunity to all suspects.

It is tempting as engineers to think that we might provide a sophisticated model of risk assessment and decision theory to proportionality judgment. But it is clear in the law that a human must make such decisions, both because such technological solutions are as yet inconceivable, but also because that entails a human who is responsible and accountable for the use of force. Even in asking whether a robot is capable of making a proportionality decision, we find ourselves looking to standards of "reasonable persons" which our robocop may not be capable of realizing, in principle. The ability of a machine learning algorithm to classify behaviors or objects does not itself constitute "reasonableness" which requires a contextual understanding of a situation.

4. Who Will be Accountable for the Automated Violence a Robot Commits?

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 the use of force when it occurs. 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.

Like law enforcement officers, a law enforcement robot system must be accountable for its use of force. At the very least this would require 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. Legal responsibility must ultimately lie with the humans who design and deploy

[48] Asaro (2012), *op. cit.*

such systems. 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.[49] Police departments will be liable to lawsuits due to the use of force by their robots. This might ensure that particular robots are kept up in proper maintenance and software updates. But could those police departments be held liable for civil rights violations if their 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. 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.

Community review boards for robots?

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 that goal are themselves highly contentious political issues. Indeed, I would argue that it is precisely because they are political that there needs 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

[49] Asaro (forthcoming) "'Hands Up! Don't Shoot!' HRI & the Use of Violent Force and Lethal Force to Serve and Protect"

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.[50] 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 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.[51] 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

---

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

[51] http://www.pcworld.com/article/209708/Is_Microsoft_Kinect_Racist.html
https://community.leapmotion.com/t/sensor-seems-to-have-trouble-with-darker-skin-color/2351
https://dgoins.wordpress.com/2015/03/21/my-kinect-told-me-i-have-dark-olive-green-skin/

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's face recognition software failing to work properly with dark faces.[52] The more recent Kinect v2 actually classifies users by the skin and hair-type, which means it could be designed to act differentially on the basis of those features.[53]

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 functionally 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.[54] 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.

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.

---

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

[53] https://dgoins.wordpress.com/2015/03/21/my-kinect-told-me-i-have-dark-olive-green-skin/

[54] http://www.cjc-online.ca/index.php/journal/article/view/2196
http://www.buzzfeed.com/syreetamcfadden/teaching-the-camera-to-see-my-skin#.id4VzgqB9

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.[55] 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, than people of color, despite it being much more likely that people of color are arrested and convicted.[56]

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 "discretionary policing."[57] That is, many of the interactions with the public that are initiated by police are at their discretion–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.

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,

---

[55] http://www.theverge.com/2014/2/19/5419854/the-minority-report-this-computer-predicts-crime-but-is-it-racist

[56] http://www.huffingtonpost.com/kim-farbota/black-crime-rates-your-st_b_8078586.html

[57] Elizabeth Joh, The New Discretionary Policing

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 decision processes often amplify and exacerbate existing racial 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[58] or from moving to certain suburbs in Atlanta.[59]

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 labeled African-American faces as "gorillas" in images.[60] 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 labeled 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.[61] 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 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

---

[58] Langdon Winner, *Autonomous Technology*, 1977.

[59] 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

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

[61] 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

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.[62] 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.[63]

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.[64] 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.[65] Given that the data sets from which machine learning of categories of suspicious persons and behavior are

---

[62] Frank Pasquale, *Black Box Society*, 2015.

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

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

[65] Joh, *op cit*.

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

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. 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.[66] 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.

[66] Asaro, P. (2000). "Transforming Society by Transforming Technology: The Science and Politics of Participatory Design," Accounting, Management and Information Technologies, Special Issue on Critical Studies of Information Practice, 10 (4), pp. 257-290.

AI Ethics in Predictive Policing: From Models of Threat to an Ethics of Care

Peter M. Asaro, *Member, IEEE*

***Index Terms*—AI Ethics, Artificial Intelligence, Design Methodology, Ethics, Machine Learning, Predictive Policing**

I. Introduction

THE adoption of data-driven organizational management–which includes big data, machine learning and artificial intelligence (AI) techniques–is growing rapidly across all sectors of the knowledge economy. There is little doubt that the collection, dissemination, analysis, and use of data in government policy formation, strategic planning, decision execution, and the daily performance of duties can improve the functioning of government and the performance of public services. This is as true for law enforcement as any other government service.

Significant concerns have been raised, however, around the use of data-driven algorithms in policing, law enforcement and judicial proceedings. This includes predictive policing–the use of historic crime data to identify individuals or geographic areas with elevated risks for future crimes, in order to target them for increased policing. Predictive policing has been controversial for multiple reasons, including questions of prejudice and precrime and effectively treating people as guilty of (future) crimes for acts they have not yet committed and may never commit. This central controversy over prejudice and precrime is amplified and exacerbated by concerns over the implicit biases contained in historic data sets, and the obvious implications for racial, gendered, ethnic, religious, class, age, disability, and other forms of discriminatory policing, as well as how it shapes the psychology and behavior of police officers.

As more bureaucratic processes are automated, there are growing concerns over the fairness, accountability, and transparency of the algorithms that are used to make consequential decisions that determine peoples' life opportunities and rights. Less discussed are the ways in which the introduction of data-centric processes and data-driven management have significant consequences on the techno-social and spatio-temporal structure of organizations [1], as well as the priorities of its management, the nature of its labor, and the quality of its results [2]. Such is the nature of contemporary technocratic governance [3]. Yet neither the increasing collection and reliance on data, nor the specific socio-technical and spatio-temporal organization of governmental institutions is determined by the technology alone, nor by the utility of data. Nor is the kind of analysis performed on that data, or the specific problems to which it is addressed, pre-determined or "natural" in any meaningful sense. Rather, there are myriad social, institutional and individual values that go into the decisions of which data to collect, when and where to collect it, how to encode it, how to assemble it in databases, how to interpret it, and how to use it to address social, institutional and individual concerns. It is those values which are the primary concern of ethics in information systems design.

This paper outlines a new ethical approach that balances the promising benefits of AI with the realities of how information technologies and AI algorithms are actually adopted, applied and used. It proposes that AI ethics should be driven by a substantive and systemic Ethics of Care, rather than by narrow Models of Threat based on utilitarian risk and threat models. While it focuses on law enforcement policies and policing practices, it hopes to contribute to the broader discussion over the ethical application of AI technologies in government policy-making and the delivery of public and commercial services more generally. The paper concludes that while data-driven AI techniques could have many socially beneficial applications, actually realizing those benefits requires careful consideration of how systems are embedded in, and shape, existing practices, beyond questions of de-biasing data. Absent such consideration, most applications are likely to have unjust, prejudicial and discriminatory consequences. This conclusion supports a proposed Ethics of Care in the application of AI, which demands moral attention to those who may be negatively impacted by the use of technology.

There is a recent and widespread excitement about the application of artificial intelligence (AI) to nearly every aspect of society–from commerce to government. AI, as a scientific research field, has long sought to develop computer programs to perform tasks that were previously thought to require human intelligence. This somewhat abstract and conditional definition has given rise to a wide array of computational techniques from logical inference to statistical machine learning that enable computers to process large and complex datasets and quickly provide useful information. Whether through traversing long chains of inference or sifting through vast amounts of data to find patterns, AI aims to provide

This work was supported in part by a Beneficial AI research grant from the Future of Life Institute.

Dr. Peter Asaro is Associate Professor in the School of Media Studies at The New School, New York, NY 10003 USA. He is also a Visiting Professor at the Munich Center for Technology in Society, Technical University of Munich, Germany, an Affiliate Scholar at the Center for Internet and Society, Stanford Law School, and an Affiliate Member at the 4TU Centre for Ethics and Technology, Twente University, The Netherlands.
(e-mail: asarop@newschool.edu)

logically sound and evidence-based insights into datasets. Insofar as these datasets accurately represent phenomena in the world, such AI techniques can potentially provide useful tools for analyzing that data and choosing intelligent actions in response to that analysis, all with far less human labor and effort. This is the traditional approach of AI, or what we might consider artificial specialized intelligence. This type of AI is essentially about creating a customized piece of software to address a complex issue or solve a specific problem by automating what would otherwise require human mental effort.[1]

Specialized AI is best seen as an extension of more traditional practices such as software engineering, IT systems design, database management and data science which deploys a range of AI techniques to automate the search for solutions to problems that currently require substantial human mental labor and skill. Much of the current excitement around AI is focused on "deep learning" machine learning techniques that use many-layered "deep" neural networks that can find complex patterns in large datasets ("big data"). Far from artificial sentience, consciousness or general intelligence, we could consider this as enthusiasm for "statistics on steroids." Commercial and governmental institutions have long used statistics to develop representations of the world that can inform future actions and policies. In this sense, the AI revolution is really a continuation, and massive acceleration, of much longer and older trends of datafication and computerization. What is new and unprecedented is the sheer volume of data, the speed at which it can now be effectively processed, the sophistication of the analysis of that data, the degree of automation and the consequent lack of direct human oversight that is possible.

As data-driven organizational management–led by big data, machine learning and AI techniques–continues to accelerate, and more processes are automated, there are growing concerns over the social and ethical implications of this transformation. Machine ethics is concerned with how autonomous systems can be imbued with ethical values. "AI ethics" considers both designing AI to explicitly recognize and solve ethical problems, and the implicit values and ethics of implementing various AI applications and making automated decisions with ethical consequences. This paper will consider the latter, implicit view which corresponds to what is sometimes called "robot ethics," to distinguish it from explicit "machine ethics" [4]. Ideally, the explicit ethics, implicit ethics, and the embedding and regulation of the system in society should all align [5].

The outputs of predictive policing algorithms clearly have ethical consequences, even if the systems under consideration do not try to design systems for explicit ethical reasoning. In the predictive policing systems under consideration, there is little or no effort to design the systems to frame their analysis or results as ethical decisions or perform ethical analyses. What is of concern to the public, and in this paper, is how well the systems are designed, and the ethical implications of introducing them into police practices.

There is a growing body of research examining the ways in which data-driven algorithms are being used in an increasing number of critical decision processes, often with little or no accountability [6, 7, 8, 9], and sometimes with little or no real understanding of how they function in the real world or why they reach the results they do in particular cases [10, 11, 12]. Consequently, there are many ways for such systems to "go wrong." Sometimes this is due to a well-intentioned but mathematically naive understanding of how such systems work. This includes the failure to understand how statistical outliers may be mishandled or misrepresented, or how historical data patterns can be self-reinforcing–such as denying credit and charging higher interest rates to poorer individuals and communities, thus systematically denying them opportunities to escape poverty. Sometimes this is due to the intended desire to transfer responsibility and blame to an automated process, and relieve human agents of their responsibility. And sometimes there may be malevolent motives behind using data in obviously discriminatory ways–such as purging voter rolls to deny eligible voters to an opposing political party. But these are ultimately "narrow" views of AI ethics, which look to improving accuracy and performance of the technology, while largely ignoring the context of use. It has also been argued that the focus of AI ethics on "solving" the bias problem is a distraction from other and more important ethical and social issues [13]. Without discounting the value of such narrow approaches, this paper will examine the importance of taking a broader ethical perspective on AI, and the problems that will not be fixed through fairness, accountability and transparency alone.

II. Two Approaches to AI Ethics

This paper aims to go beyond the ways in which data and AI algorithms might be biased or unaccountable, and consider the ethics of how AI systems are embedded in social practices. Because AI ostensively automates various forms of human reasoning, consideration and judgement, the accuracy or fairness of such processes alone do not guarantee that their use will provide just, ethical and socially desirable results. Rather, careful attention must be paid to the ways in which the implementation of such systems changes the practices of those who use them. In order to redirect attention to the bigger picture of the socio-technical embeddedness of AI when considering ethics, the paper will formulate two broad concepts of AI ethics, which will be named "Models of Threat" and an "Ethics of Care."[2] It will first outline these

[1] Some theorists have been speculating about the possibility or consequences of an artificial general intelligence (AGI) which might be able to learn with little or no direct instruction from humans, and in some sense recognize problems on its own that are in need of solution, and then adapt itself to solve them. AGI is not technologically feasible for the foreseeable future, and as such it will not be given much consideration here.

[2] Neither term is original, and each is meant to evoke traditions of thought and their general perspective, while not necessarily implying that the specific projects described were conscious of, or directly influenced by, those traditions. "Threat Modeling" has been an important methodology in cybersecurity for identifying, assessing, prioritizing and mitigating threats and vulnerabilities since at least the early 2000s [14], while "Threat Perception" has been a key concept in international relations and political psychology in assessing military threats and deterrence strategies [15]. "Ethics of Care" has been gaining popularity in medical and educational ethics since its introduction by Carol Gilligan to explain moral development in child psychology in the late 1970s

concepts in broad terms. It will then examine two illustrative cases, in the area of predictive policing, which epitomize each approach. It concludes with some observations and reflections on how to design better and more ethical AI through an Ethics of Care approach.

Perhaps the greatest ethical concerns over algorithmic decisions have been raised around the use of data-driven algorithms in policing, law enforcement and judicial proceedings. One well-researched and much discussed example from the Florida judicial system involves the use of algorithms to predict future recidivism in convicts as a basis for determining the length of their sentences.[3] Another growing application is predictive policing–the use of historic crime data to identify individuals or geographic areas with elevated risks for future crimes, in order to target them for increased policing. Predictive policing has been controversial–as it aspires to prevent crime, it also raises questions of prejudice and precrime[4] and effectively treating individuals and communities as guilty of (future) crimes for acts they have not yet committed and may never commit [21, 22]. This central controversy of prejudice and precrime is amplified and exacerbated by more general concerns over the implicit biases contained in historic data sets, and the obvious implications for racial, gendered, ethnic, religious, class, age, disability, and other forms of discriminatory policing.

Predictive policing as a term can refer to a variety of technologies and practices. The technical usage of the term usually refers to algorithmic processes for predicting locations or individuals with high probabilities of being involved in future crime, based upon historical data patterns [23]. Recent approaches utilize "big data" techniques and arguably entail forms of mass surveillance of the public [24]. However, these recent algorithmic techniques and applications have their roots in much older practices of collecting and utilizing comparative statistics (better known as CompStat) about crimes to manage large police forces, which began in New York City in 1995. While many CompStat programs utilized computer programs to calculate the statistics from crime and accident reports and arrest records and in some cases automatically generate "pin-maps" of crime activity, CompStat was really a set of data collection, analysis and management practices rather than a piece of software [25]. And CompStat has seen its share of criticism, including from former police officers [26].

Moreover, the algorithmic techniques that are increasingly being employed by police forces draw upon data that goes well beyond the digitized crime reports of the CompStat legacy, or automatically generated "heat maps" of areas of high crime activity.[5] In recent years, police departments have begun deploying and integrating large scale video surveillance systems, traffic cameras, license-plate and face recognition technologies, audio gun-shot locators, cellphone interceptors, aerial surveillance, and a host of other surveillance and data-collection technologies. As these systems become networked and produce large amounts of data, there is increased pressure to analyze, integrate and utilize this data for improving law enforcement, which leads to increased reliance on automation and algorithms for sorting and sifting through that data and translating it into policing priorities and strategies. As such, the term predictive policing can be taken to refer to a broad class of algorithmic and data-driven practices and software tools utilized by police forces. Predictive policing is also a good example of how AI might be deployed more generally, and the ethical challenges that may arise.

This paper aims to lay out a general approach to AI ethics, which is characterized here as an "Ethics of Care." It uses predictive policing, and the design of AI-based systems within it, to lay out the framework for an AI Ethics of Care. In particular it will look at two recent, but very different, implementations of data-driven interventions on youth gun violence in Chicago, Illinois, USA. Predictive policing is particularly good for this purpose for several reasons. As should be clear from the discussion above, policing is an area which gives rise to a number of critical ethical and legal issues, and has relevance not only to society at large, but to a host of other governmental functions and other industries. It is also an area that has an historical practice of data collection, and recent trials in the application of AI techniques to those practices. Further the algorithms of predictive policing embed values and make designations and decisions with implicit ethical consequences.

The Ethics of Care has a history of its own as well, and is similar in some ways to concepts in related fields, including the "Duty to Protect" in policing [28] and the "Duty of Care" in law [29]. In contrast, the Models of Threat approach construes the world and the individuals within it as risks and threats which must be managed, mitigated and eliminated. The discussion section will consider what it means to implement the Ethics of Care approach, following the examples. First the paper will give a brief sketch of each approach.

The Models of Threat approach begins from the assumption that the world can be classified into clear categories, *i.e.* threats and non-threats, and that this is the first step in choosing an appropriate action to take.[6] It focuses on capturing and processing increasing amounts and types of

and its extension by Nel Noddings into a moral theory based on interpersonal relationships of care giving and receiving in the early 1980s [16].

[3] In an analysis of 7,000 sentencing cases in Broward County, Florida over the period 2012-2013 that used the COMPAS software, journalists found similar error rates in the assessment and sentencing of white and black convicts, but diametrically opposed in their direction. White convicts were more likely to be erroneously predicted *not to* commit future crimes, while black convicts were more likely to be erroneously predicted *to* commit future crimes, resulting in shorter sentences for white convicts and longer sentences for black convicts [17].

Another study of the same dataset shows that amateur humans are able to make better predictions than the COMPAS software, using the same six factors as the software, and even better predictions can be made using just two factors–defendant's age and number of past convictions [18].

[4] "Precrime" is a science fiction concept, which first appeared in the writings of Philip K. Dick, in a novel [19] that was later turned into a major Hollywood movie [20].

[5] Such "heat maps" have become ubiquitous in the age of big data, and is even reproduced, albeit at lower resolution, on real estate websites such Trulia.com [27].

[6] This is not to say that the world, or its representation in a computational model, is necessarily discrete. One could represent the likelihood that an individual or area might present a threat or risk as a continuous variable. And while the scale and threshold for action on the basis of that variable might not be pre-determined, or determined by the system, it is expected that such metrics will influence the decisions and actions of police officers with respect to those individuals and areas—*i.e.* that the threat or risk represented by the calculation can and should result in actions.

data, and processing this to provide increasingly accurate classifiers of what constitutes a threat, and predictors of the likelihood and risk from that threat. It largely assumes that the actions that will be taken to address threats and risks are independent of the observation, collection and analysis of data. This approach also assumes that the primary values are in the accuracy, precision, fidelity, and comprehensiveness of the data model, and in the correctness of its classifications and reliability of its predictions. This approach could also be characterized as taking a narrow view, being very detail oriented, atomistic, and deeply analytic.

By contrast, the Ethics of Care approach is holistic, and takes a broad, big-picture view of the values and goals of systems design. It considers the interaction and interrelation between an action or intervention and the nature of classifying things and predicting outcomes within specific contexts. The goals and values of an Ethics of Care is to benefit everyone represented by the system as well as those who use the system, and the society as a whole. The Ethics of Care approach recognizes the complexity of social relations and socio-technical systems, including the organization using the system, and does not expect more and better data to simply solve complex social and institutional problems, but rather to provide opportunities for finding better solutions, better actions, and better policies than what are already considered.

The traditional notion of the Ethics of Care is that interpersonal relationships form the basis for normativity, and should be guided by benevolence [16]. [7] When it comes to law enforcement, we can see the Models of Threat approach seeking to better identify violations of the law, and to predict when and where violations will occur, so as to better deploy police officers to respond. It might also aim to assist police in identifying perpetrators and bringing them to justice. The Ethics of Care approach, might instead consider the factors that lead people to violate the law, and seek out new interventions that make crimes less likely and thus requiring less resources to enforce the law. It would also view the relationship between law enforcement and the community as primary and consider how any new data tool might impact that relationship.

III. A Note on "Precrime"

Beyond the practical socio-technical meanings of predictive policing, there is also a deeply troubling connotation to the term, captured in the concept of "precrime." This notion is more philosophical in nature, and draws upon our concepts of guilt, responsibility, agency, causality, and their temporality, as well as the means and ultimate aims of law enforcement in the regulation of society. The term is also mentioned extensively by nearly every press article about predictive policing, and the commercial software startup PredPol, which supplies Los Angeles and many other police departments with data analysis software, states prominently on their "About" page that they are not selling "Minority Report" technology [30]. Yet, the notion of precrime has powerful cultural meanings for good reasons beyond the popularity of sci-fi.

The basic idea of precrime stems from the idea that the goal of policing is the reduction and, ultimately, the elimination of crime altogether. While investigating crimes after they occur and responding to crimes-in-action are good, it would be even better to prevent crimes before they happen, or so this line of thinking goes. This view tends to emphasize deterrence over other key elements of criminal justice–retribution and reformation. The goal is to disrupt or dissuade criminality before it manifests. While crime prevention could focus on eliminating the means of committing crimes,[8] it more often focuses on the motives, and as such employs psychological theories of choice and sociological theories of behavior, and generally focuses on maximizing the likelihood and cost of penalties for wrongdoing by stricter enforcement and harsher penalties.[9] The temporality also becomes deeply problematic here. There is an obvious utility in preventing crimes before they occur, but our notions of individual responsibility, guilt, and punishment rest on the commission of acts–of actually doing certain things which constitute crimes–rather than imagining, desiring, or simply being psychologically pre-disposed or circumstantially inclined toward doing things which would be criminal. In some instances, planning or discussing criminal acts with others are acts that can themselves constitute a lesser crime, such as conspiracy or solicitation to commit a crime, and a failed attempt, *e.g.* to kill someone, can still constitute the crime of attempted murder even if nobody is actually hurt. But there are, and should be, different standards for citizens who have committed no crime, those in the act of committing a crime, those suspected of a crime, those convicted of a crime, and those who have served their sentences for a crime. How should law enforcement treat "those 'likely' to commit a crime"? And does the epistemic basis for that likelihood determination matter?

The classification of individuals also becomes critical here. When we say that an individual is "likely to commit a crime" is that based on their individual behavior and actions, or because of membership in a certain demographic group? "Profiling" becomes problematic in the latter case, when individuals are classified according to population-level statistics and biases. Statistics are notorious for not distinguishing correlations in data from causal reasons, and it would be unjust to treat people with suspicion for coincidental correlations when the underlying causal mechanisms for criminal behavior are absent. This kind of profiling becomes deeply problematic when it becomes prejudicial, and the

[7] According to the Internet Encyclopedia of Philosophy, "Normatively, care ethics seeks to maintain relationships by contextualizing and promoting the well-being of care-givers and care-receivers in a network of social relations. Most often defined as a practice or virtue rather than a theory as such, "care" involves maintaining the world of, and meeting the needs of, our self and others. It builds on the motivation to care for those who are dependent and vulnerable, and it is inspired by both memories of being cared for and the idealizations of self. Following in the sentimentalist tradition of moral theory, care ethics affirms the importance of caring motivation, emotion and the body in moral deliberation, as well as reasoning from particulars." [16].

[8] For instance, adding better locks to protect property, such as ignition immobilizers on cars, or making it more difficult to resell stolen goods [31]. In some cases, increasing the policing of crimes may actually have counterintuitive effects of increasing crime, according to an economic analysis of the theft of art works [32].

[9] Rarely do these approaches take into account the outright irrationality or the failure of individuals to actually think about committing crimes in rational terms. This is because cognition in the wild follows other lines of reason and risk assessment, from inflamed passions, to rational biases, to human necessity.

correlation is taken as itself constitutive of guilt, or warranting a presumption of guilt, rather than a presumption of innocence.[10]

According to the U.S. legal system, criminal liability and guilt depends upon a combination of *actus reus* (the "guilty act") and *mens rea* ("the guilty mind"). That is, one must actually commit the act for which one is held responsible, and one must have had in mind the intention, or at least the awareness, that one was doing something wrong, or should have known (as mere ignorance of the law is not a suitable defense). From this perspective, one cannot be guilty of a crime before actually committing the act, and should not be held liable for a crime not committed. And this is where precrime clashes with fundamental concepts of justice. If society, and police, act upon precrimes, and those suspected of them, in the same way as already committed crimes, then they are treating *as guilty*, or at the very least *as suspect*, those who have not yet, and not actually, committed a crime. This is a profound form of prejudice, in which judgments are made not only before relevant evidence of a criminal act can be obtained and analyzed, but before such evidence can even exist. Rather, judgement is passed on information derived from statistical inference, patterns, trends and probabilities. But a statistical likelihood of an event is neither an event nor an act.[11] And it is fundamentally unjust to treat someone as guilty of a crime they did not commit. Moreover, it is powerfully felt as an injustice when individuals and communities are treated "as if" they are guilty of doing something they have not yet, or not individually, done, based simply on their being members of a category or demographic group. Indeed, the imposition of social categories can even give rise to the new social identities [35]–and thus machine-generated categories are likely to create new types of people. This makes the creation and designation of a "criminal type" deeply problematic.

Still, there is a practical concern that law enforcement cannot ignore information about likely crimes without sacrificing their duty to prevent crime. While the scope and nature of that duty are themselves contested, this is a powerful intuition. Indeed, it is the same intuition that motivates much data-driven management. That is, if we *can* use historical data to predict future trends and events, and thus better allocate valuable resources towards fulfilling a mission or goal, then we *should* do so. While not incorrect—certainly better use of information can improve policing in many ways—if pursued without careful consideration, caution and sensitivity to its various implications and specific implementations, pursuing such intuitions blindly can quickly lead to problems. Unfortunately, the strength of this intuition and its simple logic make it an easy policy argument to make in many institutional and bureaucratic settings. One might even argue that this is the "default" policy argument in the age of data, and thus Models of Threat is the default approach to predictive policing. And it is safe to assume that without critical reflection and active awareness on the part of systems designers, something similar will be the likely default goal of most AI systems. To better understand how the design of systems can mitigate or exacerbate the problems inherent in data-driven management, we now turn to two examples of predictive policing.

IV. One City, Two Cases of Predictive Policing

The City of Chicago, Illinois has seen a spike in gun violence in recent years. The city has led the United States in the number of shootings and gun homicides, peaking with 758 total homicides and more than 4,300 shootings in 2016, and down slightly in 2017 [36]. This has led to a serious effort by the Chicago Police Department (CPD) to address this spike by focusing on the neighborhoods and individuals most likely to become involved in gun violence. A number of studies, experiments and policies have been tested and implemented in recent years. By comparing different applications of data-driven interventions occurring in the same city at the same time period, we can develop insights into the implications of data for shaping policing practices.

Two such experiments, in particular, offer a good insight into the ways in which data can be applied to address gun violence, and also into the ways that the implementation and utilization of those insights can have radically different social and ethical implications. One has been the subject of critical scrutiny by journalists and researchers, called the Strategic Subjects List. More often called the "heat list" by police officers, it was first used by CPD in 2012, and its use continues, though under a revised set of guidelines following criticism of the early uses described here. The other started in the summer of 2011 as a pilot research program implemented by the City of Chicago, and was studied the following year by University of Chicago researchers. Called One Summer, it has since been adopted as an annual program by the City of Chicago. While both started out as academic research projects, both were analyzed by outside researchers in 2012, and both utilized data to assess and identify youth who are at-risk of being involved in gun violence, in most other ways the two programs are very different.

The two projects can best be characterized as illustrative case studies, embodying two different philosophies of predictive policing, and perhaps two extremes thereof. They accordingly have very different ways of thinking about what being an "at-risk" youth means, and consequently pursue very different approaches to intervening so as to reduce that risk. More importantly, they also had very different outcomes in terms of their effectiveness in reducing gun violence and in influencing the life outcomes for those identified as "at-risk" in each program. In short, the Strategic Subjects List can be described as taking a "Models of Threat" approach to at-risk youth. That is, at-risk youth in that project are primarily viewed as threats to the community because they are at-risk,

[10] For example, if one is worried about a copycat bombing like the Boston Marathon bombing, it might make sense to flag individuals who shop for pressure cookers and backpacks. However, one should still presume there is a reasonable explanation for this rather than presuming they must be terrorists for doing so [32].

[11] Just consider gambling on horse races, which historically gave rise to modern statistics [33]. Odds-makers go to great lengths to provide accurate statistical predictions of the chances for each horse in a race. Yet, whichever horse is the favorite to win does not necessarily win–the actual outcome of the race matters. The favorite only wins about 1/3 of the time [34]. Gambling would not make sense if this were not the case–though in many games of chance it can be argued that it is mathematically irrational to place bets at all.

and interventions are targeted at increased police scrutiny and enforcement against those individuals. Whereas the One Summer program takes an "Ethics of Care" approach to at-risk youth, in which at-risk youth are given access to social services and resources aimed at reducing their risks of becoming involved in violence.[12] Like their philosophies, their outcomes were also dramatically different, despite resting on similar data-driven assessments of being "at-risk."

A. The Heat List

The Strategic Subject List (SSL) algorithm was developed as an experiment by a researcher at the Illinois Institute of Technology, and was utilized by CPD starting in 2012 and continuing until today. In its early iterations and implementations, it took data about individuals from CPD arrest records, taking into account some 48 factors, including number of arrests, convictions, drug arrests, gang affiliations, and being the victim of crimes or violence [38]. The SSL then went further, taking into account these factors for the individual's social network as determined by who was arrested together with an individual [39]. These factors were weighted and compiled into an overall SSL score from 1-500. The initial implementation contained over 398,000 individuals drawn from police arrest records, and identified 1,400 as being at "high-risk" of being involved in violence. While some 258 received the top score of 500 points, only 48% of these had previously been arrested for a gun crime, and many people on the list had never themselves been arrested, but rather were victims or were in the social networks of victims or perpetrators [39]. Many police officers reported that they were not fully informed of how the list was compiled. They assumed, or were led to believe, that everyone on the list was a perpetrator of violence and was likely to commit more violence, whereas the SSL scores combined those at-risk of being victims with those at-risk of being perpetrators in a single metric of "being involved in violence."

The practical use of the SSL list and scores was somewhat haphazard in its early years.[13] While there was no official policy regarding its use, it did feature in some CompStat reports [40], and was used by police officers in some more controversial ways. The first of these, called "custom notification," involved police officers making personal visits to high-risk individuals, informing them of their presence on the list and, further, informing them that they would be subjected to additional police scrutiny [41]. In other words, they were told that the police were "watching them" more carefully, and they should expect more police encounters. The other, and more common use of the SSL was as a "heat list" following a violent crime, in order to round-up the "usual suspects" from the list for questioning, in this case people in the vicinity of the crime who had high scores on the list. As a result, people on the list were far more likely to be detained and arrested by police, simply for being on the list. A detailed RAND study showed that the use of heat list in this way had no statistical impact on the likelihood of individuals on the list being involved in gun violence, nor on the overall gun violence in their communities [42]. It did, however, radically increase the likelihood of being arrested and convicted of a crime for those people on the list.

Further, the data and algorithm behind the SSL was not shared publicly, making it difficult to determine whether the list simply replicated long-standing racial and class discrimination. While the CPD told the Chicago Tribune that,

> "[The SSL] is not based on race, ethnicity or geographic location...We don't use it to target certain individuals other than we pay a visit to their residence to offer them services to get out of the (gang)." But a California-based group that defends civil liberties in the digital world raised concern that the arrest data that goes into it could be inherently biased against African-American and other minorities. "Until they show us the algorithm and the exhaustive factors of what goes into the algorithm, the public should be concerned about whether the program further replicates racial disparities in the criminal justice system," said Adam Schwartz, a staff attorney for the Electronic Frontier Foundation [41].

That same *Chicago Tribune* article indicates that 85% of the 2,100 shooting victims so far that year had been on the SSL, but does not indicate how they scored or whether they were all in the list of 1,400 high-risk individuals, or the longer list of 398,000 individuals included in the dataset.

Both of the main applications of the SSL, the "custom notification" warnings and using the "heat list" to bring people in for questioning, contain elements of precrime. In the warnings, there is a sense in which the police still cannot arrest an individual before a crime, but they do attempt to intimidate and threaten an individual who, in the majority of cases, has never been arrested for a violent crime. While the police do offer to "help individuals to leave gangs," it is not clear what specific services they offered, or whether those services are effective in either helping individuals get out of gangs or in avoiding future violence. Similarly, it may be an expedient tool to round up people in the area who appear on the "heat list," but it is no substitute for doing the policework of a real investigation, or following the leads from witnesses and suspects. Indeed, it may impede or undermine community-oriented policing strategies. While police may complain that witnesses, and even victims, are often unwilling to cooperate with police, these heavy-handed tactics of rounding up suspects based on data-driven lists only breaks down further the trust between communities and the police. As such, these uses of SSL actually work against confidence-building efforts by police, while offering little or no demonstrative positive results [42, 43].

Both applications also appear to engage in victim-blaming. In some cases literally so, insofar as the SSL combines victims and perpetrators in a single category of "being a party to violence" or at-risk of being "involved in violence." It makes little sense to show up at someone's door to tell them that they may be the victims of violence,[14] and less sense to threaten

[12] The slogan of the One Summer program is "Nothing Stops a Bullet Like a Job" [37].

[13] It is also worth noting that the SSL, and the data and algorithms upon which it was based, was kept private by the CPD. It was only after a long legal battle that the Chicago Sun-Times newspaper was able to force the CPD to make the SSL and its data public [39].

[14] Making someone aware of a specific threat against them would be helpful, but people are usually aware of the fact that they live in a violent neighborhood. Non-specific warnings are of little help, as has been seen with color-coded

them with increased surveillance, or to round them up for questioning after a violent crime. And detailed analysis of the effects of these practices bear out the futility of these interventions. Accordingly, this approach can best be characterized as "Models of Threat." Individuals on the SSL are seen as threats, and are themselves threatened and subjected to additional police attention, and are much more likely to be questioned and arrested. Indeed, from a crime statistics perspective, the success of a police department rests on the number of violent crimes, and many gun crimes are the result of and/or give rise to retaliation, so it makes sense to combine the victims and perpetrators of violence in a single metric. In other words, individuals likely to be involved in violence are a "threat" to the department's CompStat numbers, regardless of whether they are victims. Thus, in a Models of Threat approach, even a victim is viewed as a "threat." Yet, in any commonsense approach to violence there should be a difference in how one approaches or intervenes with an individual who is likely to be a victim from someone likely to be a perpetrator.[15] It would be difficult to argue this approach has improved policing–for instance by making police work more efficient according to its own metrics–even while it has been proven to have no effect on violent crime on either an individual or community level. And while conflating victims and perpetrators is poor data practice, it is not clear that "getting the data right" would actually improve the results of SSL. It is hoped that an AI ethic would be able to avoid such ineffectual and counterproductive applications. But to do so, it must look beyond the numbers and datasets, to understand how they are embedded in communities and policing practices.

B. Nothing Stops a Bullet Like a Job

The Ethics of Care approach offers a stark contrast to the Models of Threat. One Summer started as a pilot program in the summer of 2011 by the City of Chicago. In 2012 it became part of a controlled study (One Summer Plus) by researchers at the University of Chicago Crime Lab. The basic idea was to intervene with at-risk youth by providing them with summer jobs, for 8 weeks and 25 hours a week at minimum wage, mostly working for organizations focused on their local communities. According to the City's press release about the program, "at-risk" was defined by a combination of attending an at-risk school and a review of individual applications:

> More than 700 youth ages 14-21 were selected to participate in One Summer Plus in 2012 from an open application process available at 13 Chicago public schools located in high-violence and low-income neighborhoods. Applicants faced a number of challenges; the year before they entered the program, they had missed an average of six weeks of school and about 20 percent had been arrested [44].

As a data-driven technique, it was largely the schools which were identified through historical data. While the methodology used to identify the 13 schools is not discussed in detail, presumably it was based on the geographic location of historical incidence of violence, and the proximity of those schools to violent areas, in combination with demographic income data. But it is important to note that individual students were initially identified only in virtue of attending a designated school. The accepted applicants may have been further screened for factors such as school attendance, previous arrests, or other factors. But it is worth noting that this was not a highly sophisticated data-driven technique for identifying which individual youth were "at-risk." As far as the program was concerned, anyone living in a low-income, high-violence area was "at-risk," and more detailed or nuanced classifications were not essential to participation or effectiveness.

Researchers studying One Summer found a 51% reduction in involvement in violence-related arrests among youth who participated in the program compared to the control group that did not participate.[16] Their analysis of the data from the initial study, and of subsequent years, demonstrates that this was not simply the result of getting them off the streets for 25 hours per week, but that there were significant changes in their cognitive and behavioral approaches to school, work and becoming involved in violence [46]. Much of this was attributed to improved impulse control, learned both through their employment and through training sessions they received as part of the program. There were also economic benefits resulting from the additional income received by the participants and their families, and participants were much more likely to seek and get jobs after participating in the program.

The One Summer program provides a good illustration of an Ethics of Care approach insofar as it focuses on the contextual manifestations of violence, and seeks a means of directly intervening to change that context. Rather than focusing on the metric or individual "threat," an Ethics of Care focuses on the system. An Ethics of Care also starts from respecting people and maintains a focus on the duties and responsibilities to the individuals it deals with. By contrast, a Models of Threat approach sees people as statistics, and treats the individuals on a list as threats, whether they have done anything or not, and regardless of whether they are victims or perpetrators–thereby undermining their humanity. An Ethics of Care sees the individual as having rights and deserving of respect, and sees those at-risk as being in need of care. An Ethics of Care does not disregard data, but rather utilizes data in the service of performing a duty in a manner that respects everyone involved. And that respect extends to taking the effort and care to understand the situation from multiple perspectives, including that of citizens and working police–and how data gets used and how it relates to the lived world. Indeed, as the RAND researcher who studied the SSL says, data and AI ethics is less about sophisticated data analysis techniques and more about understanding context:

> The biggest issue for those agencies considering predictive policing is not the statistical model or tool used to make forecasts. Getting predictions that are somewhat reasonable

threat risks from the Department of Homeland Security, which do not specify any particular location or type of activity to be on the lookout for.

[15] The assumption made by researchers in doing this appears to be that there is significant overlap in the categories of victims and perpetrators. This is especially true given the cyclical nature of gun violence in Chicago, driven by rivalries and revenge killings that beget further revenge killings. Still, associating with people connected to violence might make you more likely to become a victim of violence without becoming more likely to commit violence.

[16] Subsequent research places the figure at a 43% reduction in violent arrests [45].

> in identifying where or who is at greater risk of crime is fairly easy. Instead, agencies should be most concerned about what they plan to do as a result [47].

There is a deeper lesson in this observation—the possibility of action, and the types of interventions envisioned, can strongly shape data representations, and the value of various kinds of data. While the current fashion is to collect all and any available data, in the hope that something useful might be inferable from it, there is still value in considering what actions are available to address a problem. This also means using data to find new means of acting and intervening, and better understanding the problem, rather than simply making the current means of addressing a problem more efficient. Indeed, many AI ethicists concerned about AGI worry about a hyper-efficient AGI might be so good at achieving a set goal, or maximizing a certain value, that it does so to the great detriment of other human values.[17] In the case of policing, many of the current policies and tactical goals of policing could be dangerous, unjust and counter-productive if executed with complete accuracy and efficiency. And most people would not be happy living in a society where every violation of the law was detected and punished strictly and with perfect efficiency. At least this would require rethinking many laws, policies and punishments [48]. In order to better appreciate how actions and practice could or should shape data, particularly for AI ethics, we turn now to a discussion of what the framework for AI ethics drawn from an Ethics of Care would look like.

V. AI Ethics of Care: From Data to Models to Implementation

The Ethics of Care has its own history, coming out of feminist thought. As a general normative theory, it has been criticized for failing to question what is right to do, in favor of seeking what is best to do in the circumstances. But as an approach to practical applied ethics, it has proven illuminating in areas such as educational and healthcare ethics [49, 50]. It is proposed that policing, like education and healthcare, aims to "serve and protect" the community with limited resources,[18] and as such is also a good candidate for an Ethics of Care. It is further proposed that in trying to improve the management of broad variety of governmental, non-profit and commercial organizations with data-driven techniques, AI ethics can also draw upon the Ethics of Care, as robot ethics has done [53]. In this section we look at how an Ethics of Care can be applied to data science and AI, from data collection, to data modeling, to data-driven policies and actions, drawing upon practical examples from data-driven policing.

Predictive policing, as the application of AI techniques to policing data, has its roots in much older practices of collecting crime data. Yet it also has the potential to draw upon data from other sources in increasingly networked police departments, and increasingly digitally surveilled communities. Ethical questions arise at almost every stage of data collection and analysis, from where data is collected and sensors are placed, to how data is encoded, to existing biases in segregated communities and policing practices, to the ways data is used in police management and police encounters with the public. For building a more general approach to AI ethics, it is useful to separate these problems out and identify the key ethical issues, and how AI researchers and system designers might think about and address them.

A. Data: From CompStat to Critical Data Science

Information and communication technologies (ICT) have long been central to policing. From the keeping of criminal records and crime statistics and their collection in databases, to the use of police boxes, telephones, radio dispatching and 9-1-1 emergency call centers, many of its ICT technologies have become as closely associated with policing as badges and handcuffs. Initially, these technologies were analog–paper records, photographs and inked fingerprints; dedicated police telephone boxes, and wireless radios. With the computerization of businesses and government agencies from the 1960s to 1990s, many aspects of police work also became digitized and computerized. Police patrol cars began getting computers in the early 1980s, which allowed officers to check vehicle license plates, and eventually check individuals for outstanding warrants. The transition from paper to digital records for crime reports soon led to interest in compiling

[17] Nick Bostrum's infamous paperclip maximizer which quickly and efficiently turns the world into paperclips at the expense of everyone and everything else, is an example of this.

[18] The motto of the Los Angeles Police Department, "To Protect and To Serve," was introduced in1955 following a contest at their police academy, won by Officer Joseph S. Dorobek [28]. It, and its variants, have since been adopted as the motto of numerous police departments across the United States. But what do these words really mean? The topic has been much discussed within police departments. In 1998, an Ohio police officer offered his views in *Police Magazine*,

> While what constitutes "protect" may be open to some debate, it seems to be more clear-cut than does the word "serve." It's obvious that we protect the citizens and their property from the criminal element. The word "serve" on the other hand is somewhat ambiguous. What "to serve" may mean to one law enforcement agency it may mean quite the opposite to another. "To serve" also takes on a different meaning depending upon department size. For example, I know a chief in a small village not far from the city where I work. He recently had a call to "assist the woman." We all get these types of calls, but his was to assist the woman in re-hanging her draperies! To serve? Is that what people want? A tax supported drapery service? [51]

There are two striking aspects to this passage and the article, which also seems representative of the views of many police officers, and much of the public. The first striking aspect is the extent to which "service" is framed as a question of resources. Of course, the police are public servants, as are other agents and officers of government. But they also have a specific function, and should have priorities within that function. Indeed, the rest of the article is devoted to discussing the way non-emergency calls are overloading 9-1-1 operators and keeping police from getting to real emergencies. "In many small cities, the police are the only visible and accessible arm of the local government available after 5pm and on weekends. Because of that we become the water department, the street department, the dog warden, etc.–and people begin to expect it from us." [51]

Of course, the "public" within the concept of public servant should be understood to include everyone in the community, not just "citizens" or "tax payers" or even just "law abiding" people. Police have a duty to serve everyone, including the "criminal element."

Following several court and Supreme Court decisions in the United States, there is now a legal precedent that police do not have a specific legal duty to protect, or even to enforce the law or court orders. At least in terms of having a duty to lend aid or to protect a particular individual, a police officer is not compelled by the law to intervene, put themselves at risk, or act to enforce applicable laws. The court has upheld the discretion of police to decide when and where to enforce the law or protect individuals from danger [52].

crime statistics at a local level for use in guiding the management of patrols and policing priorities. CompStat, short for Comparative Statistics, was the result. Initially adopted by the New York City police department in 1995, similar practices have been adopted across the country, especially in large urban departments.

CompStat as a mere data gathering and management practice has not been without its critics. In 2010, John Eterno and Eli Silverman, a retired New York police captain turned university professor and a criminology professor respectively, published a book-length criticism of CompStat practices in the NYPD [54]. The book argues that there was widespread mis-reporting of crimes across NYPD precincts, which took the form of downgrading the seriousness of reported crimes in an effort to show annual improvements in serious crime statistics. They argued that this systematic downgrading of crime statistics was the result of pressure from police leadership and administration. They further argued that pressures to increase police stops, especially in the era of "stop and frisk" in New York City, was highly racially discriminatory. The book caused enough controversy and embarrassment for the NYPD that the Police Commissioner ordered an independent study to review CompStat [55]. That review did indeed find serious systemic reporting errors. It did not, however, find evidence that this was the result of administrative pressure, though did not investigate that exhaustively, nor did it seriously assess systemic racism within CompStat's data collection practices.

What emerges from the investigations and reports into CompStat, from a data science and AI ethics perspective, is the susceptibility of data to political and bureaucratic pressure. While it may be convenient to assume that a given dataset offers an accurate representation of the world, this should not be taken for granted. In this case there were widespread and systematic errors in the reported data. If that data were to be used by predictive policing algorithms, those errors could have a significant impact on policing practices. And if that data is indeed racially biased, as it most likely is, it could further bias policing practices. But without an awareness of these issues, and the potential for inaccurate data or latent bias within data, the designers of those AI algorithms may be creating garbage-in-garbage-out systems, believing that they are producing quality systems (as measured by their available data). The lesson for AI ethics is to never take for granted the accuracy of given data, but to be suspicious, to seek out likely ways in which political, economic, or social pressures may have influenced historical datasets, to consider how it may be shaping current data collection practices, and to be sensitive to the ways in which new data practices may transform social practices and how that relates to the communities and individuals a system aims to care for.

With the growing popularity of AI, and increasing concerns about its impact on society, universities and professional organizations have recognized the problem and taken up the challenge of teaching ethics to the next generation of AI designers. Today, many undergraduate and graduate programs teaching AI include ethical training, but its adoption has been uneven and more could be done. Many online and professional training programs still lack critical design and ethical thinking in favor of teaching the latest techniques and tools over good design. Professional organizations including IEEE, ACM and AAAI have also led initiatives to develop ethical standards, codes of ethics, and organize a growing number of conferences and workshops on AI ethics. These are all positive developments, and it is hoped that this paper will contribute to the discussion of the ethical design of AI, especially as comes to be applied in an increasing number of socially significant and ethically consequential decisions.

While not every AI system developer can become an expert in the application domain of their techniques, the basics of critical data analysis should be taught alongside statistical techniques and machine learning techniques. In particular, system designers should be adept at recognizing the necessary characteristics of an adequate dataset, and what can and cannot be reasonably drawn from a given dataset. In many cases, only domain experts will have the kind of cultural knowledge to identify exogenous influences. This fact supports a systems design approach that includes domain experts as well as critical social scientists as members of design teams, and recognizes and respects the necessity of their expertise in shaping the ultimate system design [56].

B. Models Matter

A dataset on its own is just a collection of numbers delimited by some kind of file structure. Even decisions as to how to represent a datafield with a number–binary, integer, real, pointer, formula–can have consequences for how that data gets processed. Numbers are abstract values, which are then represented by digital numerals within computational systems. How they are numerically represented can matter. But often it is far more important how we choose to represent the world through numbers. Even when we are simply "counting" things in the world, we are also engaged in processes of classification and categorization. The data "model" that a system employs involves myriad representational choices, and seeks to serve various purposes [57].

The most obvious case in law enforcement is to characterize the law, and represent violations of the law. But there are many possible computational models of any given set of legal rules and codes, and they may not always represent the same mappings of events in the world to computational encodings. Consider the case of CompStat crime under-reporting discussed above. We could look to New York Penal Law §155.05 and §155.25 for a definition of "Petite Larceny" which is theft or withholding of property valued at less than $1000 (and not a firearm, automobile, or credit card) [58]. What if a bike has been stolen, which cost a little more than $1000 when it was new, but it is used and would likely not sell for that much, nor would an insurance company compensate its loss for more than $1000? Determining the appropriate crime requires estimating the value of the property. This is a non-trivial categorization–an auction might determine the current market value, or a bike sales expert might be able to give an appraisal, but these may not agree on the price, nor be available means for a law enforcement officer. To some extent there is discretion on the part of law enforcement, prosecutors and judges as to how to appraise and categorize such a crime–and they may take factors into account other than the strict value of the property. But once categorized, that discretionary nature tends to be erased–the crime becomes defined through

its given category, documented and entered into data collection systems. AI systems designers need to be sensitive these types of processes. Indeed, understanding data collection, and critical data representation issues should be integral to computer and information science education. Taking care in the design of AI means being able to determine what an adequate dataset is, and being able to think critically about how to define it, and what the implications of various choices of categorization are. How best to do this, in general, is a matter for further research.

C. Putting AI Into Practice

The discussion so far has focused on input–how data is structured and collected. But the presentation of data analysis, and its impact on individual and institutional practices must also be taken in account. A good example of such an issue can be seen in the use of the SSL by Chicago police. In principle, the SSL could have been used to recruit youth for the One Summer program. The choice by precincts and officers to use the list for "custom notification" and for "heat lists" following crimes are not disconnected from the design of a system like SSL. While data scientists and software engineers may wish to wash their hands of responsibility for how officers actually use their tools, they cannot. At the very least this constitutes a sort of negligence and failure to warn. Many officers were not properly or fully informed of how the list was put together, and held mistaken and problematic understandings of what it was and how it worked. The officers also lacked training, guidance and direction on how to use the system, if indeed there ever was a comprehensive plan as to how to deploy and use the system. These factors surely contributed to its misuse, and all but guaranteed its ineffectual use.

An Ethics of Care approach ought to ensure that the operators of AI systems and users of data they generate are aware of the scope and limitations of those systems. It may be too much to expect them to fully understand the computational techniques–indeed even AI experts may find the performance of certain machine learning systems inscrutable. But this does not mean that people who use these systems can be ignorant of what the system can and cannot do, how reliable it is, and what its limitations in representing the world are.

Designers also need to be aware of the context in which AI systems will be deployed and used. It should not be hard to predict what police might do with a "heat list," if one has a realistic sense of police work and the pressures operating within precincts and departments. This again points to the need for domain experts and participatory design [56]. One imagines that a police sergeant on the design team of the SSL would have pointed out the likely misuses of the system. Prototyping and testing could also help reveal such tendencies, as well as short term and long-term evaluations of the system implementation.

Transparency over the algorithms, data and practices of implementation are also necessary. While the Chicago Police Department sought to avoid embarrassment from releasing the details of the SSL, it would be impossible for independent outside researchers to evaluate its impacts–positive and negative–without access to the data and algorithms. It should not take a prolonged lawsuit from a newspaper for government agencies to share public data. Of course, as more and more commercial systems, like PredPol,[19] make the algorithms, and even the data, proprietary, they will fall under intellectual property protections. This means private companies will be processing the data, and will not be required to reveal their algorithms, or subject them to independent outside scrutiny. In some cases, private companies are even withholding crime data from the cities who produced it because they have formatted it in a database for their system and even encrypted it such that it cannot be used if the city changes to another software platform [59].

[19] PredPol is a commercial software company developing data management and predictive data systems for police departments [30].

VI. Conclusion

It is hoped that this article has shed light upon some of the central issues facing AI ethics in general and predictive policing in particular. While the use of data and AI in policing is not intrinsically or necessarily unethical, it must be done with care to avoid unjust and unethical impacts. First among these issues is that while AI ethics needs to understand the computational techniques it deploys, it also needs a critical understanding of the datasets it operates on, how data is collected, and the social organizations and the biases that those datasets may represent. This requires understanding how data practices are embedded within socio-technical systems, and not blindly analyzing data assuming that it is without bias. It is also important to understand how the use of AI tools and techniques will impact the beliefs and practices of those who engage with them. Datasets, and their computational analysis, have the power to "makeup people," and also to prejudge them according to statistical patterns and categories. Even when statistically justified, such categories, and the actions of government agents on the basis of those categories, may disrespect individual rights, human dignity, and undermine justice.

By taking an Ethics of Care approach to AI systems design and ethics, designers should have a greater awareness and respect for these issues. While any design approach is ultimately limited in its ability to mitigate all possible failures and harms, and Ethics of Care can help mitigate the most significant and widespread flaws in AI systems that will impact people's lives in consequential ways. An AI Ethics of Care has the potential to apply to areas far beyond predictive policing, and can inform many applications of AI to consequential decisions.

References

[1] Miller, Peter, & Ted O'Leary (1994). "Accounting, 'Economic Citizenship' and the Spatial Reordering of Manufacture," *Accounting, Organizations and Society*, 19(1), 15-43.

[2] Zuboff, Shoshana (1988). *In the Age of the Smart Machine: The Future of Work and Power*. Basic Books.

[3] Winner, Langdon (1977) *Autonomous Technology*, MIT Press.

[4] Asaro, Peter and Wendell Wallach (2017) "An Introduction to Machine Ethics and Robot Ethics," in Wallach, Wendell and Peter Asaro (eds.) *Machine Ethics and Robot Ethics*, The Library of Essays on the Ethics of Emerging Technologies, Routledge. Downloaded from: http://peterasaro.org/writing/WALLACH%20ASARO%20(Machine%20Ethics%20Robot%20Ethics)%20.pdf

[5] Asaro, Peter (2006). "What Should We Want from a Robot Ethic?," *International Review of Information Ethics*, **6** (12), pp. 9-16.

[6] Citron, Danielle Keats (2007) "Technological Due Process," University of Maryland Legal Studies Research Paper No. 2007-26; Washington University Law Review, Vol. 85, pp. 1249-1313, 2007. Downloaded from: https://ssrn.com/abstract=1012360

[7] Pasquale, Frank (2015) *The Black Box Society*, Harvard University Press.

[8] Selbst, Andrew and Solon Barocas (2017) "Regulating Inscrutable Systems," Presented at WeRobot 2017, downloaded from: http://www.werobot2017.com/wp-content/uploads/2017/03/Selbst-and-Barocas-Regulating-Inscrutable-Systems-1.pdf

[9] Caplan, Robyn, Joan Donovan, Lauren Hanson and Jeanna Matthews (2018) "Algorithmic Accountability: A Primer," Data & Society Technical Report, April 18, 2018. Downloaded from: https://datasociety.net/output/algorithmic-accountability-a-primer/

[10] Eubanks, Virginia (2017) *Automating Inequality: How High-tech Tools Profile, Police and Punish the Poor*, St. Martin's Press.

[11] Noble, Safiya Umoja (2018) *Algorithms of Oppression: How Search Engines Reinforce Racism*, New York University Press.

[12] O'Neil, Cathy (2017) *Weapons of Math Destruction*, Crown Random House.

[13] Powles, Julia and Hellen Nissenbaum (2018) "The Seductive Diversion of 'Solving' Bias in Artificial Intelligence," *Medium*, December 7, 2018. Downloaded from: https://medium.com/s/story/the-seductive-diversion-of-solving-bias-in-artificial-intelligence-890df5e5ef53

[14] Wikipedia, "Threat model" downloaded from: https://en.wikipedia.org/wiki/Threat_model

[15] Stein, Janice Gross (2013) "Threat Perception in International Relations," *The Oxford Handbook of Political Psychology, Second Edition*, Leonie Huddy, David O. Sears, and Jack S. Levy (eds.), Oxford University Press.

[16] Maureen Sander-Staud, "Care Ethics," *Internet Encyclopedia of Philosophy*. Downloaded from: https://www.iep.utm.edu/care-eth/

[17] Angwin, Julia, Jeff Larson, Surya Mattu and Lauren Kirchner (2016) "Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks," *ProPublica* May 23, 2016, downloaded from: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[18] Dressel, Julia and Hany Farid (2018) "The Accuracy, Fairness, and Limits of Predicting Recidivism," *Science Advances*, 17 Jan. 2018, 4(1). Downloaded from: http://advances.sciencemag.org/content/4/1/eaao5580/tab-pdf

[19] Dick, Philip K. *The Minority Report*, 1956,

[20] *Minority Report*, dir. Stephen Spielberg, 2002.

[21] Shapiro, Aaron (2017) "Reform predictive policing," *Nature*, 25 January 2017. Downloaded from: https://www.nature.com/news/reform-predictive-policing-1.21338

[22] Ferguson, Andrew Guthrie (2017) *The Rise of Big Data Policing: Surveillance, Race and the Future of Law Enforcement*, New York University Press.

[23] Kerrigan, Heather, "Data-driven Policing," Governing The States and Localities, May 2011. Downloaded from: http://www.governing.com/topics/public-justice-safety/Data-driven-Policing.html

[24] Brayne, Sarah (2017) "Big Data Surveillance: The Case of Policing," *American Sociological Review*, 82(5), pp. 977-1008.

[25] "CompStat," Wikipedia website: https://en.wikipedia.org/wiki/CompStat

[26] Rayman, Graham (2010) "NYPD Commanders Critique Comp Stat And The Reviews Aren't Good, The Village Voice, October 18, 2010. Downloaded from: https://www.villagevoice.com/2010/10/18/nypd-commanders-critique-comp-stat-and-the-reviews-arent-good/

[27] Trulia.com website: https://www.trulia.com/real_estate/Chicago-Illinois/crime/

[28] *BEAT magazine*, Los Angeles Police Department, December, 1963. Downloaded from: http://www.lapdonline.org/history_of_the_lapd/content_basic_view/1128

[29] Wikipedia, "Duty of care" downloaded from: https://en.wikipedia.org/wiki/Duty_of_care

[30] PredPol website: http://www.predpol.com/about/

[31] Josh Barro (2014) "Here's Why Stealing Cars Went Out of Fashion," *New York Times*, August 11, 2014. Downloaded from: https://www.nytimes.com/2014/08/12/upshot/heres-why-stealing-cars-went-out-of-fashion.html

[32] Frederick Chen and Rebecca Regan, "Arts and craftiness: an economic analysis of art heists," *Journal of Cultural Economics*, August 2017, 41 (3), 283-307. See summary: https://economistatkart.org/2016/05/31/why-are-there-so-many-art-thefts-and-what-can-be-done-about-them/

[33] Gabbatt, Adam (2013) "New York woman visited by police after researching pressure cookers online," *The Guardian*, August 1, 2013. Downloaded from: https://www.theguardian.com/world/2013/aug/01/new-york-police-terrorism-pressure-cooker

[34] Hacking, Ian (1975) *The Emergence of Probability: A Philosophical Study of Early Ideas About Probability, Induction and Statistical Inference*, Cambridge University Press. Second Edition, 2006.

[35] Nilsen, Rich (2012) "How Well do Horse Racing Favorites Perform?," February 12, 2012, downloaded from http://agameofskill.com/how-well-do-horse-racing-favorites-perform/

[36] Hacking, Ian, (1986) "Making Up People," in Thomas C. Heller (ed.) *Reconstructing Individualism: Autonomy, Individuality and the Self in Western Thought*, Stanford University Press, pp. 222-236.

[37] Romanyshyn, Yuliana (2017) "Chicago Homicide Rate Compared: Most Big Cities Don't Recover from Spikes Right Away," *Chicago Tribune*, September 26, 2017. Downloaded from: http://www.chicagotribune.com/news/data/ct-homicide-spikes-comparison-htmlstory.html

[38] University of Chicago, Urban Labs, One Summer Project website: https://urbanlabs.uchicago.edu/projects/one-summer-chicago-plus-nothing-stops-a-bullet-like-a-job

[39] "Strategic Subject List," *Chicago Data Portal*, website: https://data.cityofchicago.org/Public-Safety/Strategic-Subject-List/4aki-r3np

[40] Dumke, Mick and Frank Main (2017) "A Look Inside the Watch List Chicago Police Fought to Keep Secret," *Chicago Sun-Times*, May, 18, 2017. Downloaded from: https://chicago.suntimes.com/politics/what-gets-people-on-watch-list-chicago-police-fought-to-keep-secret-watchdogs/

[41] Kunichoff, Yana, and Patrick Sier (2017) "The Contradictions of Chicago Police's Secretive List," *Chicago Magazine*, August, 2017. Downloaded from: http://www.chicagomag.com/city-life/August-2017/Chicago-Police-Strategic-Subject-List/

[42] Gorner, Jeremy (2016) "With Violence Up, Chicago Police Focus on a List of Likeliest to Kill, Be Killed," *Chicago Tribune*, July 22, 2016. Downloaded from: http://www.chicagotribune.com/news/ct-chicago-police-violence-strategy-met-20160722-story.html

[43] Saundrs, Jessica, Priscillia Hunt, and John S. Hollywood (2016) "Predictions Put into Practice: A Quasi-experimental Evaluation of Chicago's Predictive Policing Pilot.," *Journal of Experimental Criminology*, September 2016, Volume 12, Issue 3, pp 347–371.

[44] Sparrow, M. K. (2016). *Handcuffed: What Holds Policing Back, and the Keys to Reform*. Brookings Institution Press.

[45] Press Release (2013) "Study: Chicago's One Summer Plus Youth Employment Program Cuts Violent Crime Arrests in Half," City of Chicago, August 6, 2013, Office of the Mayor website: https://www.cityofchicago.org/city/en/depts/mayor/press_room/press_releases/2013/august_2013/study_chicago_s_onesummerplusyouthemploymentprogramcutsviolentcr.html

[46] Sara B. Heller, (2014) "Summer jobs reduce violence among disadvantaged youth," *Science*, December 5, 2014, 346(6214), pp. 1219-1223.

[47] Hollywood, John, "CPD's 'Heat List' and the Dilemma of Predictive Policing," RAND Blog, September, 2016. Downloaded from https://www.rand.org/blog/2016/09/cpds-heat-list-and-the-dilemma-of-predictive-policing.html

[48] Hartzog, Woodrow, Gregory Conti, John Nelson and Lisa A. Shay (2015) "Inefficiently Automated Law Enforcement," *Michigan State Law Review* (2015), 1763-1796. Downloaded from: https://pdfs.semanticscholar.org/ec71/95d72b4ea51c9c6cc5d6a0e153448bbf702e.pdf

[49] "Ethics of Care" Wikipedia entry: https://en.wikipedia.org/wiki/Ethics_of_care

[50] Held, Virginia (2006) *Ethics of Care: Personal, Political and Global*. Oxford University Press, Second Edition.

[51] 1998 *Police Magazine*, Downloaded from: http://www.policemag.com/channel/patrol/articles/1998/12/to-serve-and-protect.aspx

[52] "Police Not Required to Protect," http://www.barneslawllp.com/police-not-required-protect/

[53] Van Wynsberghe, Aimee (2013). "Designing Robots for Care: Care Centered Value-Sensitive Design," *Science and Engineering Ethics*, 19(2), 407-433.

[54] Eterno, John and Eli Silverman (2010) *The Crime Numbers Game: Management by Manipulation*, CRC Press.

[55] Kelley, David N. and Sharon L. McCarthy (2013) "The Report of The Crime Reporting Review Committee to Commissioner Raymond W. Kelley Concerning CompStat Auditing," April 8, 2013 (released in July). Downloaded from: http://www.nyc.gov/html/nypd/downloads/pdf/public_information/crime_reporting_review_committee_final_report_2013.pdf

[56] Asaro, Peter (2000) "Transforming Society by Transforming Technology: The Science and Politics of Participatory Design," *Accounting, Management and Information Technologies*, 10 (4), pp. 257-290. Downloaded from: http://peterasaro.org/writing/Asaro%20PD.pdf

[57] Bowker, Geoffrey C. and Susan Leigh Star (2000) *Sorting Things Out: Classification and its Consequences*, MIT Press.

[58] New York State Penal Code, downloaded from: http://ypdcrime.com/penal.law/article155.htm?#p155.05

[59] Joh, Elizabeth (2017) "The Undue Influence of Surveillance Technology Companies on Policing," *New York University Law Review*, Forthcoming. Downloaded from: http://www.nyulawreview.org/sites/default/files/Joh-FINAL_0.pdf

Peter M. Asaro (M'10) Dr. Asaro received his PhD in the history, philosophy and sociology of science from the University of Illinois at Urbana-Champaign, Urbana, Illinois, USA, where he also earned a Master of Computer Science degree. He has held research positions at the Center for Information Technology and Policy at Princeton University, the Center for Cultural Analysis at Rutgers University, the HUMlab of Umeå University in Sweden, and the Austrian Academy of Sciences in Vienna. He has also developed technologies in the areas of virtual reality, data visualization and sonification, human-computer interaction, computer-supported cooperative work, artificial intelligence, machine learning, robot vision, and neuromorphic robotics at the National Center for Supercomputer Applications (NCSA), the Beckman Institute for Advanced Science and Technology, and Iguana Robotics, Inc., and was involved in the design of the natural language interface for the Wolfram|Alpha computational knowledge engine for Wolfram Research. He has written widely-cited papers on lethal robotics from the perspective of just war theory and human rights. Dr. Asaro co-led the IEEE Oral History of Robotics project for the IEEE Robotics and Automation Society, serves on the IEEE P7000 Standard Working Group for a Model Process for Addressing Ethical Concerns During System Design, and Chairs the Rethinking Autonomous Weapons Committee of the IEEE Global Initiative for Ethical Considerations in the Design of Autonomous Systems.