Machine Learning and Counter-Terrorism

Ethics, Efficacy, and Meaningful Human Control

Robbins, S.A.

DOI
10.4233/uuid:ad561ffb-3b28-47b3-b645-448771eddaff

Publication date
2021

Document Version
Final published version

Citation (APA)
Robbins, S. A. (2021). *Machine Learning and Counter-Terrorism: Ethics, Efficacy, and Meaningful Human Control*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:ad561ffb-3b28-47b3-b645-448771eddaff

Machine Learning & Counter-Terrorism
Ethics, Efficacy, and Meaningful Human Control

Dissertation

For the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus
prof.dr.ir. T.H.J.J. van der Hagen
chair of the Board for Doctorates
to be defended publicly on
Friday 22 January 2021 at 12:30 o'clock

by

Scott Alan ROBBINS
Master of Ethics & Technology, University of Twente, Netherlands
born in San Diego, California, USA

Composition of the doctoral committee:

Rector Magnificus,	chairperson
Prof.dr. S.R.M. Miller	Delft University of Technology, promotor
Prof.dr.ir. I.R. van de Poel	Delft University of Technology, copromotor

Independent Members:

Prof. dr. M.J. van den Hoven	Delft University of Technology
Prof. dr. S. Vallor	University of Edinburgh, Scotland
Prof. dr. N. Sharkey	University of Sheffield, United Kingdom
Prof. dr. J. Weckert	Charles Sturt University, Australia
Prof. dr.mr.ir. N. Doorn	Delft University of Technology, reserve member

Printed By: Gildeprint

For Grandpa George

In Memory of Grandma Bee

Table of Contents

Summary

Machine Learning (ML) is reaching the peak of a hype cycle. If you can think of a personal or grand societal challenge – then ML is being proposed to solve it. For example, ML is purported to be able to assist in the current global pandemic by predicting COVID-19 outbreaks and identifying carriers (see, e.g., Ardabili et al. 2020). ML can make our buildings and energy grids more efficient – helping to tackle climate change (see, e.g., Rolnick et al. 2019). ML is even used to tackle the very problem of ethics itself – creating an algorithm to solve ethical dilemmas. Humans, it is argued, are simply not smart enough to solve ethical dilemmas; however, ML can use its mass processing power to tell us the answers regarding how to be 'good', in the same way it is better at Chess or Go (Metz 2016).

States have taken notice of this new power and are attempting to use ML to solve their problems, including their security problems and, of particular importance in this thesis, the problem of countering terrorism. Counter-terrorism procedures including border checks, intelligence collection, waging war against terrorist armed forces, etc. These practices are all being 'enhanced' with ML-powered tools (Saunders et al. 2016; Kendrick 2019; Ganor 2019), including: bulk data

collection and analysis, mass surveillance, and autonomous weapons among others. This is concerning. Not because the state should not be able to use such power to enhance the services it provides. Not because AI is in principle unethical to use – like land mines or chemical weapons. This is concerning because little has been worked out regarding how to use this tool in a way that is compatible with liberal democratic values. States are in the dark about what these tools can and should do.

This thesis starts with the guiding question of how to keep meaningful human control (MHC) over ML algorithms and, more specifically, over ML algorithms used in counter-terrorism contexts. That is: how can we ensure meaningful human control over artificial intelligence in a counter-terrorism context? Of equal importance, this thesis argues that in order to achieve MHC we must avoid a technology push to use ML for the sake of innovation; rather, we must first decide what ML should be used for.

This thesis can be divided into two parts. In the first part (consisting of chapters 1-3) the focus is on a selection of (the current main) issues in the ethical debate on the use of AI by the state in security contexts in general, and in counter-terrorism contexts in particular. In the second part of the thesis (chapters 4-6), I argue against technical centered MHC, in particular explainable AI and machine ethics, and in so doing I show that ML is simply not meant to take on certain roles in security contexts, and in counter-terrorism in particular. The pragmatic goal of the second part of the thesis is to convince governments and policy makers to direct resources to ML-based solutions that are not doomed to fail, i.e. not doomed to fail for efficacy and/or ethical reasons.

Chapter 1 provides a map of the ethical issues involved in using ML as a counter-terrorism tool. Three sets of ethical issues are identified. The first set of ethical issues focus on the data used to train the algorithm. The

source, the method, and the labelling of training data all come with pitfalls that must be avoided if the resulting algorithm is to be compatible with liberal democratic principles. If we get the data wrong, then we are already headed down a path where MHC cannot be achieved. The second set of ethical issues focuses on the efficacy of the algorithm. In a counter-terrorism context, there is a great deal of focus on events that are relatively rare (e.g. terrorist attacks). The third set of ethical issues concern how to maintain MHC. ML algorithms cannot be used ethically if there are not humans who can be said to have meaningful control over them – human beings who can hold the accountability and responsibility necessary in morally salient contexts.

In chapter 2, I focus on the acceptability of data collection practices. Before an evaluation of the implementation and use of these ML algorithms, the collection of all of this data needs to be justified. Without this justification, I argue, we have already lost meaningful control over the algorithm. In this chapter I tease out ethical issues associated with bulk data collection by applying the principles of just intelligence theory. These principles are: just cause, proportionality, right intention, and proper authority.

Chapter 3 starts from the concern that the powers gained by the state in the age of bulk data collection and data science could lead to a slide into an authoritarian state. This chapter argues that transparency is an important value for preventing this slide. Transparency is an instrumental value that helps to ensure that policy and law constrain the state such that it cannot slide into authoritarianism, and further that transparency helps to assure the public that this is in fact the case.

In chapter 4 it is argued that the reasons put forward to justify the pursuit of so-called 'machine ethics' are unfounded. Machine ethics is an attempt to bypass the

problem of MHC. It is thought that if developers can endow machines with ethical reasoning capabilities, then the need for human control will be eliminated. The machines won't do anything that would require human intervention as the machines themselves are 'ethical'. In this chapter we argue that even if endowing machines with ethical reasoning capabilities was possible, the reasons put forward to do so don't stand up to scrutiny. This dashes the hopes of governments, like the US and Australia, who believe that machine ethics is what is needed in order to use ML for things like autonomous weapons systems.

Chapter 5 focuses on a principle that has been proposed by numerous scholars, institutions, and governments: explicability. I show in the chapter that making explicability a principle of 'good' AI is an attempt at MHC. Having an explanation for the output of an algorithm allows a human being to check whether or not that output was arrived at by means of criteria which violate liberal democratic values. I argue that although explicability has good motives, it fails to solve the ethical issue it is purportedly there to solve, namely providing MHC. This is for two reasons: the first being the property of 'requiring an explanation' and the second being the 'irrelevance of AI' if the conditions for a decision are already known.

Chapter 6 pivots from the topic of explicability and introduces the concept of envelopment for establishing boundaries within which the use of a particular ML-powered application is acceptable. Enveloping an algorithm, in short, is to constrain it. It aims at allowing the algorithm to achieve the desired output given limited capacities. I suggest that ML algorithms should be enveloped, and, furthermore this envelopment requires transparency about other aspects of ML algorithms. These aspects are the: training data, expected inputs, possible outputs, function or functions of the algorithm (i.e. the

purpose) and, boundaries within which the algorithm is expected to operate. Transparency on these factors provides the knowledge needed to make decisions about where, when, and how, these algorithms can be used.

Returning to the original question, how can we ensure meaningful human control over artificial intelligence in a counter-terrorism context? I conclude that we must meet these necessary conditions: the algorithm is trained using data collected in compliance with liberal democratic values; the tasks delegated to AI are in part legitimized by the people they are supposed to protect, through active transparency; we do not delegate outputs to AI which require explanations; and that we have created an envelope for the algorithm.

In the conclusion I hypothesize that the kinds of outputs requiring explanations are evaluative ones that have morally significant consequences, e.g., have potentially harmful consequences if realized - such as surveilling, detaining, or shooting dead persons suspected of being terrorists who might not be terrorists. I suggest that part of keeping MHC over machines means restricting machines to outputs that do not amount to value judgments. Machines that can make decisions based on opaque considerations should not be telling humans what decisions morally ought to be made and, therefore, how the world morally ought to be. Delegating these judgments of moral value to machines is a reduction of human control over our most important sphere of decision-making. Not only would we be losing control over specific decisions in specific contexts, but we would be losing control over moral decision-making in particular. Outsourcing moral decision-making in general, and certainly in counter-terrorism contexts, to AI-powered machines, will forever result in a loss of MHC.

Samenvatting

Machine Learning (ML) bereikt het hoogtepunt van de hype-cyclus. Als je kunt denken aan een persoonlijke of grote maatschappelijke uitdaging - dan wordt ML voorgesteld om deze op te lossen. ML zou bijvoorbeeld kunnen helpen bij de huidige wereldwijde pandemie door het voorspellen van COVID-19-uitbraken en het identificeren van dragers (zie bijvoorbeeld Ardabili et al. 2020). ML kan onze gebouwen en energienetwerken efficiënter maken - en zo helpen de klimaatverandering aan te pakken (zie bijvoorbeeld Rolnick et al. 2019). ML wordt zelfs gebruikt om het probleem van de ethiek zelf aan te pakken - het creëren van een algoritme om ethische dilemma's op te lossen. Mensen, zo wordt betoogd, zijn eenvoudigweg niet slim genoeg om ethische dilemma's op te lossen; ML kan echter zijn massale verwerkingskracht gebruiken om ons de antwoorden te geven over hoe we 'goed' kunnen zijn, net zoals het beter is in schaken of het spel Go (Metz 2016).

Staten hebben kennis genomen van deze nieuwe macht en proberen ML te gebruiken om hun problemen op te lossen, waaronder hun veiligheidsproblemen en, van bijzonder belang in deze dissertatie, het probleem van de terrorismebestrijding. Procedures voor

terrorismebestrijding, zoals grenscontroles, het verzamelen van inlichtingen en het voeren van oorlog tegen terroristische strijdkrachten worden allemaal "verbeterd" met instrumenten die op het gebied van ML werken (Saunders et al. 2016; Kendrick 2019; Ganor 2019). Hieronder vallen bijvoorbeeld: het verzamelen en analyseren van gegevens in bulk, massabewaking, en autonome wapens. Dit is verontrustend. Niet omdat de staat dit soort machtsmiddelen niet zou mogen gebruiken om zijn diensten te verbeteren. Niet omdat artificiële intelligentie (AI) in principe onethisch is om te gebruiken - zoals landmijnen of chemische wapens. Het gebruik van ML instrumenten in terrorisme bestrijding is verontrustend omdat er weinig is uitgewerkt over hoe dit instrument te gebruiken is op een manier die verenigbaar is met liberale democratische waarden. Staten tasten in het duister over wat deze instrumenten kunnen en moeten doen.

Deze dissertatie begint met de leidende vraag hoe men zinvolle menselijke controle ('meaningful human control', MHC) kan houden over ML-algoritmen en meer specifiek over ML-algoritmen die worden gebruikt in de context van terrorismebestrijding. Dat wil zeggen: *hoe kunnen we zorgen voor zinvolle menselijke controle over artificiële intelligentie in de context van terrorismebestrijding?* Van even groot belang is dat deze dissertatie stelt dat we, om MHC te bereiken, een technologische drang naar het gebruik van ML omwille van de innovatie moeten vermijden; in plaats daarvan moeten we eerst beslissen waarvoor ML gebruikt dient te worden.

Deze dissertatie kan worden onderverdeeld in twee delen. In het eerste deel (bestaande uit de hoofdstukken 1-3) ligt de focus op een selectie van (de huidige belangrijkste) kwesties in het ethische debat over het gebruik van AI door de staat in veiligheidscontexten in het algemeen en in terrorismebestrijdingscontexten in het bijzonder. In het tweede deel van het proefschrift

(hoofdstukken 4-6) pleit ik tegen technisch gecentreerde MHC, in het bijzonder verklaarbare AI ('explainable AI') en machine-ethiek ('machine ethics'), en laat ik daarmee zien dat ML eenvoudigweg niet bedoeld is om bepaalde rollen in veiligheidscontexten, en in terrorismebestrijding in het bijzonder, op zich te nemen. Het pragmatische doel van het tweede deel van het proefschrift is om overheden en beleidsmakers te overtuigen om middelen te sturen naar op ML gebaseerde oplossingen die niet gedoemd zijn te mislukken, d.w.z. niet gedoemd zijn te mislukken om redenen van effectiviteit en/of ethiek.

Hoofdstuk 1 geeft een overzicht van de ethische aspecten van het gebruik van ML als instrument voor terrorismebestrijding. Er worden drie sets van ethische kwesties geïdentificeerd. De eerste reeks ethische kwesties is gericht op de gegevens die worden gebruikt om het algoritme te trainen. De bron, de methode en het labelen van de trainingsdata hebben allemaal valkuilen die moeten worden vermeden, wil het resulterende algoritme compatibel zijn met de liberaal-democratische beginselen. Als de data niet klopt, dan is MHC al niet meer bereikbaar. De tweede reeks ethische kwesties richt zich op de effectiviteit van het algoritme. In de context van terrorismebestrijding is er veel aandacht voor gebeurtenissen die relatief zeldzaam zijn (bijvoorbeeld terroristische aanslagen). De derde reeks ethische kwesties betreft de vraag hoe het MHC in stand kan worden gehouden. ML-algoritmen kunnen niet ethisch worden toegepast, als er geen mensen zijn die zinvolle controle hebben over de algoritmen - mensen die de verantwoording en verantwoordelijkheid kunnen dragen die nodig is in een morele context.

In hoofdstuk 2 concentreer ik me op de aanvaardbaarheid van gegevensverzamelingspraktijken. Voordat een evaluatie van de implementatie en het gebruik van deze ML-algoritmen

kan worden gemaakt, moet het verzamelen van al deze gegevens worden verantwoord. Zonder deze rechtvaardiging, zo argumenteer ik, hebben we al een zinvolle controle over het algoritme verloren. In dit hoofdstuk daag ik ethische kwesties in verband met het verzamelen van bulkdata uit door de principes van de rechtvaardige intelligentietheorie toe te passen. Deze principes zijn: *rechtvaardige oorzaak, proportionaliteit, juiste intentie* en *juiste autoriteit*.

Hoofdstuk 3 gaat uit van de zorg dat de bevoegdheden die de staat in het tijdperk van de bulkdataverzameling en de datawetenschap heeft verworven, kunnen leiden tot een afglijden naar een autoritaire staat. Dit hoofdstuk stelt dat transparantie een belangrijke waarde is om deze verschuiving te voorkomen. Transparantie is een instrumentele waarde die ertoe bijdraagt dat het beleid en de wetgeving de staat zodanig inperken dat deze niet kan afglijden naar een autoritaire staat. Daarbij draagt transparantie ertoe bij dat het publiek er zeker van is dat beleid en wetgeving dit ook daadwerkelijk bewerkstellingen.

In hoofdstuk 4 wordt betoogd dat de redenen die worden aangevoerd om het streven naar de zogenaamde 'machine-ethiek' te rechtvaardigen, ongegrond zijn. De machine-ethiek is een poging om het probleem van MHC te omzeilen. Men denkt dat als ontwikkelaars machines kunnen uitrusten met ethische redeneringen, de noodzaak van menselijke controle zal worden geëlimineerd. De machines zullen niets doen wat menselijke tussenkomst zou vereisen, aangezien de machines zelf 'ethisch' zijn. In dit hoofdstuk beargumenteren we dat zelfs als het mogelijk is om machines te voorzien van ethische redeneervermogen, de redenen die worden aangevoerd om dit te doen niet standhouden bij nader onderzoek. Dit doet afbreuk aan de hoop van regeringen, zoals de VS en Australië, die geloven

dat machine-ethiek nodig is om ML te gebruiken voor zaken als autonome wapensystemen.

Hoofdstuk 5 richt zich op een principe dat door vele wetenschappers, instellingen en regeringen is voorgesteld: verklaarbaarheid. Ik laat in het hoofdstuk zien dat het maken van verklaarbaarheid tot een principe van 'goede' AI een poging is tot MHC. Het hebben van een verklaring voor de output van een algoritme stelt een mens in staat om te controleren of die output al dan niet tot stand is gekomen aan de hand van criteria die in strijd zijn met liberale democratische waarden. Ik stel dat de verklaarbaarheid weliswaar goede motieven heeft, maar geen oplossing biedt voor het ethische probleem dat het zou moeten oplossen, namelijk het verstrekken van MHC. Dit is om twee redenen: de eerste is de eigenschap van 'uitleg vragen' en de tweede is de 'irrelevantie van AI' als de voorwaarden voor een beslissing al bekend zijn.

Hoofdstuk 6 draait om het onderwerp verklaarbaarheid en introduceert het begrip *envelopment* voor het vaststellen van grenzen waarbinnen het gebruik van een bepaalde ML-aangedreven toepassing acceptabel is. Envelopment van een algoritme betekent, kortom, het inperken van het algoritme. Het is erop gericht het algoritme in staat te stellen de gewenste output te bereiken binnen gegeven beperkte capaciteiten. Ik stel voor dat envelopment van ML-algoritmen toegepast dient te worden, en bovendien dat deze envelopment transparantie vereist over de andere aspecten van ML-algoritmen. Deze aspecten zijn de: trainingsgegevens, verwachte input, mogelijke output, functie of functies van het algoritme (d.w.z. het doel) en de grenzen waarbinnen het algoritme naar verwachting zal functioneren. Transparantie over deze factoren levert de kennis op die nodig is om beslissingen te nemen over waar, wanneer en hoe deze algoritmen kunnen worden gebruikt.

Terugkomend op de oorspronkelijke vraag: *hoe kunnen we zorgen voor zinvolle menselijke controle over artificiële intelligentie in de context van terrorismebestrijding?* Ik concludeer dat we aan deze noodzakelijke voorwaarden moeten voldoen: het algoritme wordt getraind met behulp van gegevens die zijn verzameld in overeenstemming met liberale democratische waarden; de taken die aan AI worden gedelegeerd, worden deels gelegitimeerd door de mensen die ze geacht worden te beschermen, door middel van actieve transparantie; we delegeren geen output aan AI die uitleg behoeft; en we hebben voor envelopment van het algoritme gezorgd.

In de conclusie veronderstel ik dat de soorten uitkomsten die uitleg vereisen evaluatieve uitkomsten zijn. Die hebben moreel significante gevolgen, bijvoorbeeld, als ze potentieel schadelijke gevolgen hebben als ze worden gerealiseerd - zoals het in kaart brengen, vasthouden of doodschieten van personen die ervan verdacht worden terroristen te zijn, maar die misschien geen terroristen zijn. Ik stel voor dat een deel van het houden van MHC over machines betekent dat machines worden beperkt tot uitkomsten die geen waardeoordelen opleveren. Machines die beslissingen kunnen nemen op basis van ondoorzichtige overwegingen zouden de mensen niet moeten vertellen welke beslissingen moreel gezien genomen zouden moeten worden en dus hoe de wereld *moreel gezien* zou moeten zijn. Het delegeren van deze oordelen van morele waarde aan machines is een vermindering van de menselijke controle over ons belangrijkste gebied van de besluitvorming. We zouden niet alleen de controle verliezen over specifieke beslissingen in specifieke contexten, maar we zouden ook de controle verliezen over morele besluitvorming in het bijzonder. Het uitbesteden van morele besluitvorming in het algemeen, en zeker in antiterrorismecontexten, aan AI-aangedreven machines, zal voor altijd resulteren in een verlies van MHC.

Introduction

The Age of Machine Learning

Machine Learning (ML) is reaching the peak of a hype cycle. If you can think of a personal or grand societal challenge – then machine learning is being proposed to solve it. For example, ML is purported to be able to assist in the current global pandemic by predicting COVID-19 outbreaks and identifying carriers (see, e.g., Ardabili et al. 2020). ML can make our buildings and energy grids more efficient – helping to tackle climate change (see, e.g., Rolnick et al. 2019). ML is even used to tackle the very problem of ethics itself – creating an algorithm to solve ethical dilemmas. Humans, it is argued, are simply not smart enough to solve ethical dilemmas. ML, however, can use its mass processing power to tell us the answers regarding how to be 'good' – in the same way it is better at Chess or Go (Metz 2016). What each of these examples has in common is a push to use ML rather than the identification of ML as being the best possible tool to assist in the problem solving of a particular issue.

Before going further into the debate on the ethics of ML it is first important to clarify some definitions. First, what is ML and how does it relate to Artificial Intelligence (AI)? AI is an umbrella term that covers a variety of methodologies to create the appearance of

intelligence. An algorithm that completes a maze by simply following the wall, either right or to the end left (called the 'wall follower') qualifies as AI because the algorithm will be able to handle many mazes that are given to it. This makes AI go above and beyond mere automation – which would be an algorithm programmed to make a series of turns that solves only the maze at hand. If given another maze, the algorithm demonstrating automation would fail. The 'wall follower' algorithm appears intelligent because it can solve a new maze that even the programmer had never seen. It is this appearance of intelligence – that is, outputs that result from inputs based on an environment that appear to be directed at achieving the 'best' outcome – that results in an algorithm classified as AI (Russell and Norvig 1995, chap. 1). That is, AI is a practice that aims to result in machines that 'act rationally'.

There are other definitions of AI centered on, for example, the aim to result in machines that 'act like humans.' This is the definition favored by Alan Turing and is the purpose of the Turing Test (Turing 1950; Russell and Norvig 1995, 2–3). I do not wish to enter into debates about the 'true' purpose of AI here; I merely want to clarify how I use the term throughout this dissertation.[1] To that end, while I agree with the definition favored by Stuart and Norvig, I do not use AI to simply mean machines that act rationally. I am specifically talking about machine learning (ML), which is one particular methodology for creating AI. To be sure, I use AI and ML interchangeably unless otherwise noted.

ML is the sub-field of AI which allows algorithms to change how they produce outputs based on previous inputs. Programmers do not give explicit rules or instructions on how to process a particular input into an output; rather,

[1] Although I fail to see the purpose of getting machines to act like humans. We have billions of humans. I hope that computer scientists and roboticists are doing something more interesting with AI.

based on a large amount of sample inputs (i.e., training data) the algorithm is able to learn a reliable way of processing inputs in the future. This ability to generate outputs without human given rules creates the pressing need for meaningful human control (MHC). Symbolic AI (or good old-fashioned AI) is simply a complex set of given rules and instructions provided by humans. This, to be sure, can complicate ascriptions of responsibility due to its complexity; however, ML puts the problem of MHC in the spotlight. It is also ML that has put AI into the limelight in the last decade. ML has realized some of the power promised by AI.

States have taken notice of this new power – and are attempting to use it to solve their problems, including their security problems and, of particular importance in this thesis, the problem of countering terrorism. Counter-terrorism procedures including border checks, intelligence collection, waging war against terrorist armed forces, etc. These practices are all being 'enhanced' with ML-powered tools (Saunders et al. 2016; Kendrick 2019; Ganor 2019), including: bulk data collection and analysis, mass surveillance, autonomous weapons and so on. This is concerning. Not because the state should not be able to use such power to enhance the services it provides. Not because AI is in principle unethical to use – like land mines or chemical weapons. This is concerning because little has been worked out regarding how to use this tool in a way compatible with liberal democratic values. States are in the dark about what these tools can and should do.

To be fair there is a growing body of literature that discusses the negative impacts that AI can have on individuals and groups. The range of ethical and societal concerns includes but is not limited to: reinforcing and/or exacerbating societal biases; making it even more difficult to ascribe responsibility and accountability to

human beings when things go wrong (also referred to as the responsibility gap); and a concern for negative impacts on democracy and democratic process. We know that these problems exist – and they will occur when the state uses AI as well – but we know little about how to overcome them.

There is also a growing body of literature, from academics, policy makers, and civil society organizations, with ideas on how to mitigate the ethical concerns raised thus far: making algorithms 'explainable'; designing machines which can recognize and act on ethically salient features (i.e. machine ethics); principles or rules of conduct to follow in the development and use of AI, to name a few. There are so many lists of principles that meta-research has been done on them (Fjeld et al. 2020).[2] However, these ideas are not translating into specific requirements that will help liberal democratic states to use ML-powered tools in a manner consistent with their avowed values, whether in counter-terrorism contexts or elsewhere.

What many of the proposed solutions have in common is that they take as a given that AI can be used for any purpose. Do you want to use AI to predict who will become a terrorist? Sure – as long as you have done it in line with the principles given – or made it 'explicable' – or you have added an ethics module to its code. This is putting the cart before the horse. Like any tool, we must first understand what the tool should be used for if we want to keep it compatible with liberal democratic values. We don't allow CCTV cameras in children's bedrooms and then try to come up with rules for how to ensure privacy might be maintained. We step back and say that CCTV cameras are not appropriate for use in children's bedrooms.

[2] Examples of organizations with lists of principles for AI include: The EU's High Level Expert Group on AI (2019), The Future of Life Institute (2017), Google (2018), Partnership on AI (2019), etc.

This brings us to the crux of this thesis. This thesis starts with the guiding question of how to keep MHC over ML algorithms and, more specifically, over ML algorithms used in counter-terrorism contexts. That is: *how can we ensure meaningful human control over artificial intelligence in a counter-terrorism context?* Of equal importance, this thesis argues that before we can achieve MHC we must first decide what ML should be used for.

Meaningful Human Control over Algorithms

The more the state relies upon machine outputs to handle increasingly important decisions in general, and in relation to security in particular, the more important it is that humans have meaningful control over those machines. Decisions made in counter-terrorism, such as decisions about who to kill in a drone attack, who to lock up, who to target for intrusive surveillance, etc. obviously have profound implications for liberal democracies and their citizens. It should not be the case that a person is labeled as a terrorist due to an algorithm over which no human can claim to have some kind of control. Of course, much hinges on what 'meaningful control' amounts to. To date, there are diverse proposals to ensure that machines are under (our) meaningful control (Santoni de Sio and van den Hoven 2018; Heikoop et al. 2019; Chengeta 2016; Crootof 2016; Horowitz and Scharre 2015; Mecacci and Santoni de Sio 2020; Robbins 2020). I suggest dividing these two proposals into two groups: Technology-centered MHC and Human-Centered MHC. The former attempts to establish control by putting technical requirements on the algorithm, e.g. the algorithm must be explainable. The latter focuses on where the human being sits in the process and what that human being has the power to do, e.g. shut down the process.

Technology-Centered Meaningful Human Control

Technology-centered MHC's two most notable initiatives are machine ethics and explainable AI. Machine ethics is the

research program with the goal of endowing machines with moral reasoning capabilities. The resulting machines have been named Artificial Moral Agents (AMAs) or Moral Machines (Wallach and Allen 2010). Specific implementations of AMAs range from trying to develop machines that 'read' literature to help them to understand human values (Riedl and Harrison 2016) to devising 'ethical subroutines' for machines based on the moral philosophy of the author's favorite philosopher (Anderson and Anderson 2007).

The intuition guiding the development of AMAs is that the world is so complex that it will be impossible to dictate what a machine's output should be in every context that the machine will inevitably face. If the machine can be guided by human values instilled in said machine, then the reasoning goes that the outputs of the machine will be aligned with human values and, in counter-terrorism contexts, potentially the relevant liberal democratic principles. In this reading, MHC is about designing machines in a way such that there won't need to be direct human control over their outputs. Technology-centered MHC, then, is about designing machines that won't have outputs that humans would need to intervene to prevent the occurrence of.

While those who design AMAs may have good intentions, the reality is that AMAs cause more problems than they solve. Chapter 4 of this thesis argues against the development of AMAs, and concludes that: "considering that no critical or unique operational function appears to be gained through the endowment of ethical reasoning capabilities into robots…[we should] place a moratorium on the commercialization of robots claiming to have ethical reasoning skills." AMAs would exacerbate the problem of MHC by tasking the machines themselves with the most meaningful decisions possible – ethical decisions. In counter-terrorism contexts, ethical decisions are

heightened and encompass questions such as: who to shoot dead, who to detain, who to intrusively surveil and so on. Machine ethics, in my view, is more about giving up on the problem of MHC than solving it.

In another technology-centered MHC approach, one could require algorithms to have explanations of their outputs. This would enable humans to ensure that the considerations used to come up with a particular output do not violate societal values. For example, if a machine's output is to deny a particular person a visa and the explanation that comes along with it includes the consideration that the person is 'of middle-eastern appearance', then a human being could reject the machine's output as being biased.

In chapter 6 I argue that this solution to MHC faces a Catch-22: "If [ML] is being used for a decision requiring an explanation then it must be explicable AI and a human must be able to check that the considerations used are acceptable, but if we already know which considerations should be used for a decision, then we don't need [ML]." The crux of the argument is that once we have made explicit the considerations that should be used to make a particular decision then there is no need to use ML as good old-fashioned AI would be possible. Making AI explicable is a fascinating engineering problem that could be put to good use in generating considerations that should be used to make certain decisions; however, this benefit is epistemic rather than normative. It may increase our ability to make better decisions after the fact – but does nothing to realize MHC for any particular decision.

The final example of a proposal for technology centered MHC is 'Track and Trace' proposed by Filippo Santoni de Sio and Jeroen van den Hoven (2018). The idea is to put two conditions that must be met in order to realize MHC. The first is a tracking condition which is about the outputs of machines being verifiably responsive to human

moral reasons. That is, if a morally salient feature were to be added to a particular situation which would change the decision a human would make, then the algorithm should change its output in the same way. For example, an algorithm might determine that someone is a terrorist due to the fact that they had consistently downloaded additions of the Islamic State's *Inspire* magazine and searched for ways for terrorists to evade surveillance. A human might come to the same conclusion. However, if the morally salient feature "the person in question is doing a PhD thesis on counter-terrorism" then a human would change their mind. The algorithm should also be responsive to this morally salient feature and not classify the person as a terrorist.

The second condition is a tracing condition which is about the output of a machine being traceable to a human being. That human being should be in an epistemic situation whereby they understand both the machine's capabilities and possible impacts on the world, and that others may have legitimate moral reactions towards them. This does not, however, mean that they understand how the machine arrived at a particular output. In chapter 6, I present my argument for 'enveloping' AI-powered algorithms which goes towards realizing this condition ('envelop' will be discussed further below).

Human-Centered Meaningful Human Control

Human-centered MHC is about the placement of a human being into an otherwise machine-based process to ensure that MHC is achieved. The guiding question is: what role can we assign to a human being to ensure that a machine's output is under her control? The two main ways to ensure human-centered MHC are to place a human in-the-loop or a human on-the-loop. The former places a human being in the process in a manner that ensures it is the human who approves or denies each output. The latter places a human as an observer of the process so that they can intervene

if necessary. Important to note is that without human intervention the output will still occur. To illustrate the consequences of the absence of a human-in-the-loop or a human-on-the-loop (i.e. humans are out-of-the-loop) consider a predator drone with face recognition technology that is programmed to kill Osama bin Laden. If there is no human in or on the loop then the drone, once programmed and activated, would detect, track and kill the individual determined by the drone to be bin Laden without the possibility of further human intervention.

Putting a human in the loop seems to be the most immediately obvious solution to MHC. Simply put, an output of an AI-powered machine needs to be confirmed by a human operator before the process is complete and the consequences of the output are realized. At first glance, this appears to be a simple way to overcome problems of responsibility and accountability. However, much research has been done to show that this solution does not work as advertised. In particular, three human biases stand in the way of this solution working properly: automation bias, assimilation bias, and confirmation bias.

Automation bias has been described in detail by Professor Cummings. It "occurs when a human decision-maker disregards or does not search for contradictory information in light of a computer-generated solution which is accepted as correct" (Cummings 2012). In a nutshell, automated systems increase a human's correct action when the automated system is correct (as opposed to those without the automated system); however, when the automated system is incorrect, humans are less likely to come up with the correct action compared to humans without an automated system – even when both have access to the same evidence.[3]

[3] For a detailed study involving pilots and automated decisions see Skitka et al. (1999).

Another form of bias, assimilation bias, is often brought up in the discussion regarding autonomous weapons systems. Humans in a certain context who have been provided an output will likely place that output in a coherent narrative and attribute intentions to people that follow that narrative. Professor Noel Sharkey uses the example of a human operator determining whether or not to go ahead with a lethal drone strike in a counter-terrorist operation, when the algorithm outputs that a strike should occur. The human operator sees people loading items onto a truck and must decide on initiating a strike. Although the items being loaded were mundane bales of hay, the output of the algorithm (i.e. that a strike should occur), in addition to the narrative context they are in (i.e. searching for terrorist activity), causes the human operator to believe that the people are loading rifles into the truck. In this context, the human operator is looking for dangerous behavior – and finds it whether it exists or not (Sharkey 2014).

The third form of bias, confirmation bias, occurs when humans seek out evidence that confirms their prior beliefs or the hypothesis on hand (Sharkey 2014). If a machine tells you that a person has a gun, then you look for evidence that a gun is on that person. This is in contrast to looking for evidence that contradicts the machine's output. If one is only looking to confirm the machine's output, then it is more likely that they will find evidence that supports the output and miss evidence that disconfirms it.

These three biases make the human-in-the-loop approach fall short of *meaningful* human control. Humans-in-the-loop as a form of human-centered MHC simply puts human beings in an incredibly difficult spot whereby they must overcome powerful biases to assert their control.

Alternatively, another approach to human-centered MHC exists, namely, humans-on-the-loop. Putting a human on the

loop is close to letting the machine operate autonomously. The only difference is that a human being is put in charge of monitoring the machine and given the power to stop the machine or overturn an output. First, this approach falls victim to the biases mentioned above. However, this solution also has a further problem which can be highlighted by autonomous vehicles, that is a decrease in the human on the loop's preparedness to take over control.

Let's say you have owned an autonomous vehicle for 6 months. When operating in autonomous mode the vehicle has been, so far, 100% reliable. In autonomous mode, you, as the human, are on the loop. That is, you observe the actions of the vehicle and have the power at any time to take control. Given its 100% reliability to date, how closely will you be paying attention to every move the car makes? How will you tell the difference between a swerve which is necessary to prevent a crash, and a swerve that is life-threatening to the human in the vehicle? As it turns out, the situational awareness of humans in autonomous vehicles (in autonomous mode) has shown to be decreased– meaning that the time it takes someone to take over control will be greater than the time it takes a critical event to occur (de Winter et al. 2014).

All this is not to say that we shouldn't have humans in or on the loop. Humans should be involved in monitoring automated systems. However, as we have just seen, humans being in the loop does not guarantee that the control they are able to exert will be meaningful. Accordingly, it will be disingenuous to assign moral and legal responsibility to those humans after a critical failure.[4]

In studying the various approaches discussed above, all with the intention to mitigate or prevent ethical issues from occurring, a new concern is revealed, that of

[4] This makes Tesla's policy of assigning responsibility to the person at the wheel when a crash happens unacceptable.

technological solutionism (or optimism). It seems to be taken for granted that AI should be used for all sorts of tasks and decisions, that there is no task or decision that will not, or should not, be delegated to AI. The MHC project, then, is about how to do all of this 'ethically'. But, let us take a step back for a moment and ask: what if there are some tasks and decisions that should not be delegated to AI? What if it is (already) unacceptable to delegate to AI the task of deciding whether to: shoot someone dead or not, classify people as terrorists, or label people in airports as 'suspicious'? Before figuring out how to design and implement AI in a 'responsible' way, we must decide what AI can responsibly be used for. In this thesis, I take a step towards understanding what kinds of tasks and decisions AI should, and should not, be used for. My contribution with this thesis to the ongoing debate about responsible/trustworthy/ethical AI is therefore about asking a different kind of question, namely, what are the boundaries within which we should be using AI in the context of counter-terrorism.

Thesis Overview

This thesis can be divided into two parts. In the first part (consisting of chapters 1-3) the focus is on a selection of (the main) issues in the ethical debate on the use of AI by the state in security contexts in general, and in counter-terrorism contexts in particular. In the second part of the thesis (chapters 4-6), I argue against technical centred MHC, in particular explainable AI and machine ethics, and in so doing I show that ML is simply not meant to take on certain roles in security contexts, and in counter-terrorism in particular. The pragmatic goal of the second part of the thesis is to convince governments and policy makers to direct resources to ML-based solutions that are not doomed to fail, i.e. not doomed to fail for efficacy and/or ethical reasons.

In greater detail, Chapter 1 provides a map of the ethical issues involved in using ML as a counter-terrorism tool. Three sets of ethical issues are identified. The first set of ethical issues focus on the data used to train the algorithm. The source, the method, and the labelling of training data all come with pitfalls that must be avoided if the resulting algorithm is to be compatible with liberal democratic principles. If we get the data wrong, then we are already headed down a path where MHC cannot be achieved.

The second set of ethical issues focuses on the efficacy of the algorithm. In a counter-terrorism context, there is a great deal of focus on events that are relatively rare (e.g. terrorist attacks). This makes the amount of training data relatively small – possibly too small to properly train an algorithm. Furthermore, one must prioritize between precision (the number of true positives divided by the number of true + false positives) and recall (the number of true positives divided by the number of true positives + false negatives). So, for example, let's say we have 1000 people buy a plane ticket to go to a particular destination. Twenty of these people are terrorists. When the algorithm classifies someone as a terrorist correctly it is called a 'true positive.' When the algorithm classifies someone as a terrorist incorrectly it is called a 'false positive.' If the algorithm incorrectly classifies someone as a non-terrorist it is called a 'false negative' and if the algorithm correctly classifies someone as a non-terrorist then it is a 'true negative.' An algorithm could achieve a high accuracy rate by simply categorizing everyone as 'not a terrorist'. While the algorithm would get it wrong 20 times it would still have a 98% accuracy rate since 980 people out of 1000 were correctly identified as not being terrorists. Nevertheless, this would be a terrible situation since none of the 20 terrorists were identified. This example helps to show why accuracy is not the best

value to look at in examples in which we are, so to speak, looking for a needle in a haystack (as opposed to, for instance, chicken-sexing (roughly 50% are males/50% females, but it is hard to tell males from females and vice-versa). This situation gives us a 0% recall rate and a 0% precision rate. Ideally we want 100% of both, but that is incredibly unlikely.

Alternatively, let us consider if we focus on achieving precision in an algorithm and it ends up classifying 10 people as terrorists, one of which is a false positive. This gives the algorithm a 90% precision rate as 90% of the time when it classifies someone as a terrorist, it is correct. However, this algorithm would have only a 45% recall rate – as it only identified 9 out of the total 20 terrorists we're searching for. Now, tweaking the algorithm to make the recall rate higher will generally lower the precision of the algorithm (e.g. more non-terrorists will be classified as terrorists). Thus, these two ML values are in tension with one another and both have ethical significance since it is a problem to arrest an innocent person but also a problem to fail to arrest a guilty person.

Equally important to consider is that efficacy, in the counter-terrorism context, is not necessarily a fixed target – meaning that ML algorithms trained on data from the past may be unreliable for predictions. Terrorist groups know that they are being surveilled and actively change tactics to evade the intelligence community. Choosing the right goal for an algorithm, therefore, is extremely important. No theory of MHC will make up for an algorithm which simply doesn't work. Choosing an achievable goal for an algorithm, therefore, is important.

The third set of ethical issues concern the maintenance of MHC. In a counter-terrorism context the outputs of ML algorithms will often have ethically salient consequences – e.g. placing someone on a terrorist watch list or a no-

fly list. These classifications will have a significant impact on a human being's life. Someone should be held accountable and responsible for these classifications. For this to occur, they must be able to exercise control over the output. ML algorithms can make this type of control difficult due to the opacity of the considerations which led to a particular output or classification. In short, Chapter 1 is a discussion of the ethical pitfalls awaiting the attempt to provide machine learning-based solutions to some, if not many, counter-terrorism problems.

Chapter 2 focuses in depth on the data collection aspect of ML, even an adequate quantum of reliable data. To develop ML algorithms for any context requires a significant amount of data. The algorithm must take examples from the past in order to classify novel inputs. For example, algorithms designed to catch terrorists must be trained on data from people already known to be terrorists. Using ML for countering terrorism, therefore, requires the state to collect large amounts of data – much of which is not directly associated with known targets. This means they are collecting data on innocent, and often uninformed, civilians. By collecting as much data as possible the state increases their chances of collecting data from persons who are, in fact, terrorists (thereby increasing the recall rate). This gives the state more data to train algorithms with – but also more chances for algorithms already trained to find the terrorists they were trained to find.

An important part of chapter 2 is to pay tribute to the acceptability of data collection practices. Before an evaluation of the implementation and use of these ML algorithms, the collection of all of this data needs to be justified. Without this justification, I argue, we have already lost meaningful control over the algorithm. The collection of bulk data has received significant attention since the Snowden revelations in 2013. Bulk data is

distinguished from targeted data in so far as bulk data will mostly include data associated with innocent civilians. Privacy activists, civil society organizations, and citizens increasingly argue that the government simply should not be able to collect such data. To counter this, the intelligence community has argued that much of the data in question wasn't actually 'collected' since it was filtered out. Moreover, such data collection can be justified, according to the intelligence community, as it is necessary to fight against threats like terrorism.

Chapter 2 teases out ethical issues associated with bulk data collection by applying the principles of just intelligence theory. These principles are: *just cause*, *proportionality*, *right intention*, and *proper authority*. The application of just cause, for example, demands that one ask what it is, exactly, that is being evaluated. In this chapter, I argue that the object to be evaluated is not the method of collecting data in bulk as a whole; rather, the object to be evaluated are the filters which result in data being collected in bulk. The filters specify the groups of people who will have their data collected. For example, a filter could specify that all data that has been encrypted will be collected. This provides us with something to evaluate; namely, evaluating whether or not there is just cause to collect data from people who use encryption. An analysis may conclude that without further attributes added to the filter (e.g. collect data that has been encrypted AND comes from Syria) the filter does not meet the just cause principle. I argue that there is just cause for the bulk collection of data when the filter description refers to a group for which there is evidence indicating said group is engaged in terrorist activity directed at the nation-state that is collecting the bulk data.

The application of the principle of proportionality to bulk data collection to determine whether or not it is disproportionate requires empirical evidence that is not yet available. Quantifying the harms of bulk data collection is not easy. While we know that there is some evidence that government surveillance – including bulk data collection – causes a 'chilling' effect amongst citizens (a chilling effect occurs when people alter their behavior due to real or perceived surveillance), the true extent of this effect – and the resulting consequences for the functioning of the state - are yet to be understood. Furthermore, the government's use of third parties for data collection (e.g. Google, Facebook) could cause economic consequences due to citizens' concern about being surveilled. Further work must be done to understand the consequences of this kind of intelligence collection for the application of a principle of proportionality.

Applying the principle of right intention to bulk data collection forces us to consider the duration of the storage of data collected. An important distinction to consider here is that between initiation of a filter and the duration for which the filter is in place. Applying a principle of right intention merely at the initiation of a filter leaves out the fact that the data collected will be stored for, possibly, a indefinite period. Furthermore, the filter will continue to collect data long after it was put in place. I argue that the data collected by a specific filter should be tied to the cause used to justify that filter. When that cause no longer exists the data should be deleted.

Finally, chapter 2 looks at the principle of proper authority. As just intelligence theory is based on just war theory the only proper authority is the state. However, reliance on third-party technology companies complicates this matter. The Snowden revelations revealed the use of programs like PRISM which gave the intelligence

community direct access to the data on company servers to enable it to find terrorists. This blurs the line between a technology company collecting data for doing business and collecting data for countering terrorism. I argue here that if data is considered necessary for national security then the government should not contract the collection of that data to third parties.

Chapter 3 (written together with Adam Henschke)[5] starts from the concern that the powers gained by the state in the age of bulk data collection and data science could lead to a slide into an authoritarian state. This chapter argues that transparency is an important value for preventing this slide. We argue that transparency is an instrumental value that helps to ensure that policy and law constrain the state such that it cannot slide into authoritarianism, and further that transparency helps to assure the public that this is in fact the case. When transparency is not realized then there is no possibility of a public debate regarding what policy and law are necessary to properly constrain the state. Furthermore, there is no way to stop the public from believing that the state is collecting data in accordance with liberal democratic principles. In this chapter, Adam and I recommend concrete actions that should be taken to help realize the instrumental value of transparency.

This chapter is crucial given the need for legitimacy of tactics used in the context of counter-terrorism, a legitimacy that is given by the citizens of the state. Because it is the state using AI for counter terrorism, there must be a degree of consent by the citizens of that state to use it. It is not just the state that needs MHC over these algorithms, but the citizens of the state must be assured that they have MHC over what the state is doing.

[5] Adam and I divided this work evenly so we each did 50% of the writing. My writing comes out most clearly in sections 3.3, 3.4, and 3.5 – though we each had a hand in every section.

This requires a level of transparency to ensure that democratic debate can occur.

In chapter 4, together with Aimee van Wynsberghe[6], it is argued that the reasons put forward to justify the pursuit of so-called 'machine ethics' are unfounded. Machine ethics is an attempt to bypass the problem of MHC. It is thought that if developers can endow machines with ethical reasoning capabilities, then the need for human control will be eliminated. The machines won't do anything that would require human intervention as the machines themselves are 'ethical'.[7] The literature on machine ethics gives seven reasons justifying the project to provide machines with ethical reasoning capabilities. These reasons are: (1) machines will inevitably be delegated roles requiring such reasoning, (2) moral reasoning is necessary to prevent harm to humans, (3) machines are so complex that novel situations will result in unpredictable actions – and moral reasoning is required to make sure those actions are ethical, (4) moral reasoning is required for the public to trust them, (5) moral reasoning is necessary if machines are to prevent their being used by humans for immoral acts (e.g. telling a robot to kill a baby), (6) machines will be better than humans at moral reasoning, and (7) through trying to equip machines with moral reasoning capabilities we will gain a better understanding of human morality. Although I am skeptical that a machine can be equipped with 'ethical reasoning capabilities', in this chapter we argue that even if it was possible, these reasons don't stand up to scrutiny. This dashes the hopes of governments like the US and Australia who hope that machine ethics is what is

[6] Aimee and I divided this work evenly so we each wrote 50%. My writing comes out most clearly in sections 4.3.1, 4.3.4, and 4.3.6 – though we each had a hand in every section.
[7] I do not agree with Machine Ethicists' premise that a machine can ever be 'ethical' in the sense of moral agency.

needed for them to use machine learning for things like autonomous weapons systems.

Chapter 5 focuses on a principle that has been proposed by numerous scholars, institutions, and governments: explicability. I show in the chapter that making explicability a principle of 'good' AI is an attempt at MHC. Having an explanation for the output of an algorithm allows a human being to check whether or not that output was arrived at by means of criteria which violate liberal democratic values. For example, if an algorithm labelled someone a terrorist and provided an explanation which showed that this output was primarily based on the fact that the person had a beard and was Muslim then this output would be biased against a religious group. Furthermore, the output would be based on an irrelevant consideration (the fact that the person has a beard). This output, then, should be discarded – and a human being equipped with such an explanation would have the information necessary to exercise that control over the use of the outputs of the algorithm.

I argue that although explicability has good motives, it fails to solve the ethical issue it is purportedly there to solve, namely providing MHC. This is for two reasons: the first being the property of 'requiring an explanation' and the second being the irrelevance of AI if the conditions for a decision are already known. For the former – the property of requiring an explanation - I argue, that an explanation belongs to the action or output and not the process which leads to the action or output. This means that the process leading to an output must be accompanied by an explanation in virtue of the fact that the output is the kind of output that requires an explanation – NOT in virtue of the process itself. Labelling someone a terrorist requires a justifying explanation regardless of how that label was decided upon given that such a label will have real world consequences,

e.g. a person labelled a terrorist will have his or her autonomy restricted in the form of being on a no-fly list. We could not, for example, simply claim that the process leading to that label has been reliable in the past – whether that process is a particular algorithm or an individual intelligence analyst. Both processes should be required to provide a justifying explanation because the result of this label restricts an individual's autonomy. Accordingly, it is not AI then that needs to be explicable, but whatever process is delegated the task of labeling people as terrorists.

Second, for an explanation to equip some human with the information necessary to establish meaningful control, that human must already have the considerations that are both relevant and compatible with liberal democratic values that would justify labeling someone a terrorist. However, if that human already has such information then the use of a ML algorithm, I argue, is no longer necessary.

Chapter 6 pivots from the topic of explicability and introduces the concept of envelopment for establishing boundaries within which the use of a particular ML-powered application is acceptable. Enveloping an algorithm, in short, is to constrain it. It aims at allowing the algorithm to achieve the desired output given limited capacities. I suggest that ML algorithms should be enveloped, and, furthermore this envelopment requires transparency about other aspects of ML algorithms. These aspects are the: training data, expected inputs, possible outputs, function or functions of the algorithm (i.e. the purpose) and, boundaries within which the algorithm is expected to operate. Transparency on these factors provides the knowledge needed to make decisions about where, when, and how, these algorithms can be used. In other words, this knowledge gives the information needed to envelop these algorithms.

To remind the reader, this thesis set out to address a specific problem in the AI ethics debate, namely, *how can we ensure meaningful human control over artificial intelligence in a counter-terrorism context?* The chapters of part 2 (chapters 4-6) point in the direction of a solution to this question. Chapter 4 argues that there is no good reason for endowing machines with moral reasoning capabilities. Therefore, trying to overcome the problem of MHC by creating machines that can 'control' themselves is problematic at best and dangerous at worst. Chapter 5 claims that outputs that 'require an explanation' should not be delegated to machines. Thus, MHC is about understanding the kinds of outputs that are suited for ML algorithms. That is, if we don't choose the right outputs for machines then we have failed to achieve a necessary condition for MHC. Chapter 6 adds another dimension to the concept of MHC – namely, that in order to achieve MHC of AI, such AI-powered machines should be enveloped. In order to envelop AI-powered machines, greater transparency pertaining to development, verification, and execution of the algorithm is needed. Thus, MHC is a noteworthy concept, one that resists technological determinism, anchoring the consequences of ML to the humans developing and using the technology. Unfortunately, the ways in which MHC have been articulated to date, presents developers with insurmountable problems (e.g. automation bias, etc). If, however, one explores a divergent articulation of MHC, one that places envelopment at the core and directs human control not only for use of the algorithm but for choosing ML as a solution to a particular problem in the first place, then one is able to avoid many of these pitfalls.

Chapter 1.

Machine Learning, National Security, and Ethics[8]

1.1 Introduction

Machine learning (ML) is being promoted as a balm to fix nearly all of the world's problems. It purportedly will help us find love, find and cure diseases, flag fake news and propaganda, defeat hackers, and, especially relevant to our purposes here, prevent terrorism. In many domains (e.g., intelligence, policing, healthcare, etc.), the problem was having a lack of data. Now there is simply too much data – so much that it cannot reasonably be processed by human beings in a time frame that would be helpful. The speed and predictive power of ML is a natural fit for this problem.

[8] A version of this is to be published as: Robbins, S. "Machine Learning, National Security, and Ethics" (forthcoming).In Clarke, M., Henschke, A., & Legrand, T. (eds), Palgrave Handbook of National Security. Palgrave.

The context of bulk data collection for national security presents us with serious concerns – both regarding its ethical acceptability as well as its efficacy. ML has been proposed to help with both of these. Regarding its acceptability in liberal democracies, advocates of ML hope to reduce violations of privacy by reducing the amount of data that actual humans will look at. The idea being that algorithms going through your data, as opposed to humans, does not constitute a privacy violation. Regarding its efficacy, advocates of ML claim that these algorithms will enhance the intelligence community's ability to: collect relevant data, detect suspicious behavior, and predict terrorist attacks.

If true, this would indeed be a dramatic achievement. However, many pitfalls await national security practitioners if and when they adopt ML algorithms for the collection and analysis of bulk data. The first set of pitfalls concerns the *data* that is used to train these algorithms. This data can come from the surveillance apparatus (including bulk data collection), governmental and public records, and third-party technology companies. As will be shown below, these sources face concerns of dual-use of the data, the legitimacy of the source itself, and how securely it holds the data.

Moreover, we must ask why is it justified to use a particular dataset to train an algorithm for surveillance purposes? We will see that while it is possible to justify some bulk collection, this justification precludes the use of ML for such collection. Finally, for the data to be useful for ML, it must be labeled. This labeling can reinforce the biases of the past and cause disparate impacts for different groups of people.

Once the ML algorithms are trained, another set of issues arise concerning its *efficacy*; namely, how can we be sure that these algorithms will be better than the processes that preceded them? Biased or unrepresentative training

data can lead to ineffective algorithms. An algorithm that is biased towards labeling those with darker skin as suspicious – as happened recently with face recognition technology - could miss suspicious people with lighter skin – reducing its effectiveness by increasing false-positives and false-negatives. Further, depending upon what the ML algorithm is used to look for, there may not be enough data to train it. Algorithms looking for so-called 'lone wolf' terrorists, for example, would be challenging to train because there are very few such attacks – and even fewer data with which to train an algorithm.

Even after we have an effective, well-trained algorithm using responsibly acquired training data, there is the critical issue of *meaningful human control*. That is, having some human who will be responsible and accountable for the algorithm's decisions. In morally salient contexts like surveillance and policing, clarity regarding responsibility and accountability are paramount. ML algorithms are opaque concerning the reasoning that led it to deliver a particular output. This makes it even more difficult for humans to have meaningfully control over them. The last section addresses the solutions currently on offer for this problem of human control.

1.2 Machine Learning for Mass Surveillance

Before we examine the many pitfalls awaiting those developing and implementing machine learning solutions for surveillance, we need a snapshot of what ML is already used for in the national security context. Below is a non-exhaustive list of ML solutions for a variety of tasks. Many are already being used in the field as I write this.

1.2.1 Facial Recognition / Smart Surveillance Cameras

So-called 'smart surveillance,' like CCTV cameras before it, is on its way to becoming ubiquitous. We hear about many of its applications in China; however, tools for

smart surveillance are also marketed to police departments and intelligence agencies in the West.[9] These new ML enhanced cameras can, in real-time, detect suspicious behavior, the faces of criminals, license plates, falls, violence, etc. In a nutshell, these smart surveillance cameras can perform object recognition in real-time – tracking object movements and classify behavior.

Governments are turning to such technologies to fight crime, counter-terrorism, and to facilitate 'public opinion guidance.' The latter phrase is used by the Chinese government to monitor dissent. A New York Times investigation revealed that 18 governments are using technology developed to conduct mass surveillance in China (Mozur et al. 2019). It is not merely institutions of the state using these cameras. Smart cameras are being used by private entities as well as individual consumers to monitor their businesses and homes.

1.2.2 Voice Recognition

Most of us are now able to command our smartphones via digital assistants like Siri and Google Assistant. Our smartphones (and other smart devices like Amazon's Alexa or Google Home) can take commands because ML algorithms are trained to understand speech. Now speech recognition is being taken a step further. Not only can it convert speech to text (which allows it to understand commands), but it can identify who is speaking. Amazon's Alexa, for example, can give personalized responses based on who is talking to it. Alexa can do this because ML can match audio input to a specific person.

The possibilities of using this for surveillance are huge. Searching on a database of recorded phone calls by a specific voice would be extremely helpful. When a known terrorist's voice is 'heard' by the ML algorithm, an alarm

[9] See the advertisement for Gorilla's IVAR smart surveillance system for a look at some of the functionality being offered: https://www.gorilla-technology.com/IVAR.

could alert the authorities. The European Union (EU) recently completed a project called the Speaker Identification Integrated Project (SIIP). The project hoped to identify criminals and terrorists from their voices. The project took audio from social media posts, internet websites, and lawfully[10] intercepted communications data. In addition to being able to identify people based on voice, the system "identify gender, age, language, and accent, and detect voice cloning" (Interpol 2017).

The National Security Agency (NSA) has a voice identification system that can identify terrorists, criminals, potential whistleblowers, etc. Analysts describe the system as "Google for voice." The NSA "could use keywords and "selectors" to search, read, and index recordings that would have otherwise required an infinite number of human listeners to listen to them" (Kofman 2018).

1.2.3 Attack Prediction

Using machine learning to predict the details of a terrorist attack before it occurs would help national and local officials divert resources, prepare, and hopefully prevent the attack. Using data from past terrorist attacks or communications data leading up to such attacks could allow algorithms to tell us something about a future attack. Police are using similar algorithms to predict crime hotspots and criminals with what is now called 'predictive policing.' While there is much debate surrounding the ethics and efficacy regarding the use of such algorithms, adoption amongst police departments is increasing.[11] Examples include PredPol and Palantir. PredPol advertises their use of machine learning in the

[10] Note that lawfully does not mean ethically.
[11] Over 60 American police departments use predictive policing algorithms ("How data-driven policing threatens human freedom" 2018)

products directed towards police departments while Palantir is secretive about their specific techniques.

Others have used sentiment analysis on social networking sites like Twitter to predict who potential terrorists are. For example, Azizan & Aziz (2017) use the sentiments towards words associated with terrorism like "bomb," "ISIS," "Muslim," etc. to update an overall sentiment towards terrorism, which is based on historical sentiments to those same keywords. The hope is that those with a more positive overall sentiment towards terrorism (reduced to the keywords they chose as being associated with terrorism) are more likely to commit terrorist attacks.

1.2.4 Financial Fraud

When many banks receive a deposit above a particular threshold, say over $10,000 in cash, they are required to submit a Suspicious Activity Report (SAR). There is widespread agreement that an amount of $10,000 or above in cash is suspicious. Authorities hope that these SARs will help combat the financing and profit of illegal activities. Of course, how to classify a transaction or a set of transactions as 'suspicious' is, in many cases, subjective. It differs from one institution to another, and those trying to evade scrutiny change their behavior to avoid detection. As criminal financial activity becomes more sophisticated, the ability of financial institutions to detect it decreases. The patterns may be simply too intricate for humans to see.

Algorithms have long aided financial institutions with the identification of suspicious activity; however, until recently, these algorithms used "good old fashioned AI" (GOFAI) in the form of if-then statements (see above). GOFAI, in its simplest form, is a decision tree that automates previously decided upon human reasoning. A human could follow the same decision tree to reach the same output in a GOFAI application. ML, on the other hand, is being deployed to detect suspicious patterns that a human

could not. The Israeli company ThetaRay, for example, is one of many companies now offering machine learning products to fight terrorists and organized crime financing and money laundering. They use unsupervised machine learning to identify behavior that was previously unknown to be associated with money laundering. This is significant as machine learning is generally used to detect patterns based on past known events. For example, there may be a specific set of transactions that are associated with money laundering. Machine learning might detect that pattern and then notify banks when similar patterns are caught in the future. ThetaRay goes beyond this to highlight previously unknown patterns of activity that deviate from the norm – which may detect new money laundering techniques before they are successful.

1.2.5 Propaganda on Social Media

The spread of propaganda on social media platforms is receiving a lot of attention from the media (see, e.g., Overly 2017). Terrorist organizations use it to spread misinformation and recruit new members. Mass shooters are posting live feeds of themselves killing people indiscriminately. Platforms like YouTube and Facebook are consistently criticized for not taking preventing the spread of these messages. They are now turning to ML to take down propaganda as soon as it is uploaded.

The British home office is using ML to detect Islamic State (ISIS) propaganda videos. They claim that the algorithm detects 94% (recall) with 99.995% accuracy (precision). The algorithm was trained on over 1000 propaganda videos (Home Office 2018). Facebook claims that 99% of Al Qaeda and ISIS posts are detected by machine learning before a user flags the post (Horwitz 2017). Google says that it will use its "most advanced machine learning research to train new 'content classifiers' to help us more quickly identify and remove extremist and

terrorism-related content" ("Four Steps We're Taking Today to Fight Terrorism Online" 2017).

1.3 Societal Implications of the Training Data

The first set of issues one faces concerning any machine learning application stems from the data that is used to train the algorithm. The use of machine learning in highly morally significant contexts – national security applications like policing, counter-terrorism, and intelligence – amplifies these issues. When a local police department purchases ML-powered software, enabling some form of 'smart' surveillance, they should understand that there are significant concerns with that software (Henschke 2017). The source of that data may be using the data for other purposes (e.g., targeted advertising), may be illegitimate when it comes to state surveillance, and may not have the security citizens often expect when it comes to sensitive personal data.

1.3.1 The Source

If you were to find out that your local police department was using facial recognition to run smart surveillance in your neighborhood, you would be justifiably concerned. However, if you found out that the algorithm running facial recognition was trained on billions of images scraped from the internet – including social media, personal websites, and YouTube – you might be angry. And anger is what many felt when they heard about ClearView AI – a company that trained their ML algorithm on precisely that – and works with law enforcement around the world (Hill 2020).

Three general issues can arise due to the source of training data for machine learning algorithms: legitimacy, dual-use, and data security. Legitimacy refers to the concern that it may be unjustified to receive and use data from a particular source. It would be unjustified, for

example, if police officers went door to door and scanned your photo albums to collect data for their new machine learning algorithms because private citizens' homes are an illegitimate source for such data. Even if one were to commit a crime resulting in a search warrant for their home – it might be illegitimate to collect and use anything other than evidence relevant to the crime at hand.

How can we determine if a source is legitimate? There are a few variables that can help. The first variable is what the purpose is of the collected data. If the data is connected to a specific algorithm with a particular function that we would (if asked) consent to, then the source may be legitimate. For example, if police had evidence that a serial killer or mega-terrorist (known to have perpetrated the 9/11 attack) worked in your office building, then work emails, browsing histories, and other network data may be collected justifiably. That is, your office's network activity is a legitimate source when it comes to the specific function of finding a specific serial killer or terrorist. In this view, the purposes of the surveillance are necessary aspects of their justification (Henschke 2017, pp. 245–251).

The problem so far is that the algorithms that will be trained by the data collected have no such specific purpose. There is no particular person that the authorities are trying to find. Furthermore, the uses of these algorithms are not tethered to a narrow set of 'serious' crimes. This unclarity makes it challenging to know in advance whether or not using data to build a specific algorithm is justified. Facial recognition algorithms, for example, may be used to ticket jaywalkers. This purpose is not severe enough to justify the collection and use of people's pictures.

The second variable that helps determine source legitimacy is who actually will be collecting the data from this source. It is one thing, for example, for government

actors (e.g., police or intelligence agencies) to collect and use this data and quite another for a private company contracting with the government to do so. This is because government actors in liberal democratic countries have accountability mechanisms and oversight whereby abuse of powers are uncovered and perpetrators punished.

1.3.1.1 Dual Use

Dual-use technologies commonly refer to technologies that have both military and civilian purposes. Think of a nuclear power plant. The energy it provides is a benefit to society (although there are downsides regarding the storage of nuclear waste). However, a nuclear power plant can also be used to create materials central to nuclear weapons. Data collection for ML algorithms is touted as providing safety (a benefit). However, that same data can be used by the state for harmful purposes.

The widespread public acceptance of 'smart' technologies has given the state potential access to more and new data. Digital assistances like Alexa and Google Home provide them with voice data. Smart cameras like Amazon's Ring provide video. Online picture storage applications allow them to obtain a database of valuable image data. Smartwatches and smartphones provide location data (i.e., GPS). Social networks can get data on our social relations, political beliefs, and mood. While these devices may provide users some convenience in their daily lives, they open up the possibility for widespread surveillance by the state. There have been numerous cases of a country requesting or demanding data from these devices. For example, Amazon was ordered to turn over audio recordings collected by an Amazon Echo device for evidence in a double murder case (Cuthbertson 2018).

As I write this (May 2020), there is a raging debate regarding COVID-19 contact tracing smartphone applications. While this technology has been touted as a solution to re-opening countries from their prolonged

lockdown,[12] there are fears that the state could also use this technology to track dissident groups and political rivals. To many, the concern that state institutions are collecting and storing your personal data may seem to be rooted in paranoia and would only be a real concern for those who are criminals. However, in many places, such data is used to detect who dissidents and political opponents are. Furthermore, there is a concern that as people come to realize that the authorities could potentially access the data collected about them by technology companies, they will modify their behavior because they are embarrassed by some of their behavior or are scared of being put on watch lists by the government. This 'chilling effect' has been little studied; however, it has been shown to cause self-censorship (Gao 2015b). Those fearful that the government is watching are less likely to express dissent – dissent which is crucial to a functioning democracy.

1.3.1.2 Legitimacy

Due to the widespread surveillance capitalism apparatus set up by technology companies, it is cause for little surprise that security agencies, such as the police, make requests to technology companies for the personal data of their customers to assist them in their criminal, including counter-terrorism, investigations. Amazon Alexa data has been used to investigate a murder, police are working with Ring to gain access to their doorbell cameras, and Edward Snowden revealed the degree to which intelligence agencies have access to the databases of Apple, Google, Facebook, Microsoft, and others.

While a specific request under warrant (and, therefore, with judicial approval) for particular data about a serious crime may well be justified, other practices that involve private companies in the collection of

[12] Though there is little evidence that these contact tracing apps have any benefit (Vogelstein and Knight 2020)

intelligence for security agencies may not be. At any rate, there is a concern regarding the legitimacy of technology companies, in particular, conducting activities generally reserved for the state. For instance, citizens have not consented to a private company like Google monitoring their online behavior for evidence of criminal activity. In the surveillance and intelligence literature, it is argued that there is a need for an organization to comply with a principle of 'legitimate authority' if it is to conduct a surveillance or intelligence program (see, e.g., Macnish 2014). Another approach looks at whether the use of surveillance products is legitimate based on the justifications for that surveillance (Henschke 2017, pp. 253–254). On this approach, as mentioned above, the police and intelligence agencies may be justified in requesting specific data from a technology company that may pertain to a particular investigation. For example, no one would object to the police requesting and obtaining call records from a telecommunications company regarding a suspect who is believed to have committed murder. Data collected by technology companies, therefore, should not be completely off-limits to the authorities. The line is crossed when technology companies are forced to collect and retain data that would not have been collected for purposes other than government surveillance. In other words, if a company does not need a dataset to operate their business (and would likely delete it), then they should be able to do that.

1.3.1.3 Security

Finally, as private technology companies are increasingly used as sources, collectors, and processors of data, there is an increasing risk of sensitive data being accessed by unauthorized malicious actors. It is one thing for the state's security agencies to safely and securely collect, process, and store this data with stringent oversight mechanisms in place to prevent unauthorized access and misuse. Quite another for a technology company to be

entrusted to do so, given that a private company's primary institutional aim is to maximize profits and, therefore, its safety, security, and oversight mechanisms might be far less stringent. For example, a third-party technology company was used to facilitate the monitoring of people and cars (using facial recognition and license plate recognition) passing through certain United States borders. This company was breached, resulting in 100,000 records being stolen. The Customs and Border Patrol (CBP) agency is quoted as saying that 'none of their systems were compromised.' In effect, they have passed on the responsibility of safeguarding sensitive data collected for use by the state to a private technology company. Their quote makes it appear that the CPB does not see the breach as a failure, given the standards required of their systems.

Situations like the CPB example above leave the state in a dilemma. Either they are abnegating responsibility, thereby admitting that their use of a third party company is illegitimate, or they must accept responsibility for the breaches and, therefore, be blamed for the breaches of security on the part of the third party technology companies they use. This is important because these kinds of arrangements in which the state, in effect, outsources data collection, storage, and processing to third party technology companies are increasing. Both the National Security Agency (NSA) and the Central Intelligence Agency (CIA) in the US have agreements to store and process data using Amazon Web Services (AWS). The US military, in its defense strategy document (Department of Defense 2018), says it will "will enhance partnerships with U.S. industry to align civilian AI leadership with defense challenges." It is essential that the security of the data involved in these partnerships is strong – and that the state accepts responsibility for said data.

1.3.2 Labeling

Once training data has been sourced and collected responsibly, there are still issues stemming from the labeling of that data. ML algorithms are taught what a cat looks like in a similar way children are taught. They are given many examples of pictures or live images of cats and told that the photo or video contains a cat. After many such instances, the algorithm can reliably classify an image or video as having a cat or not.

Unfortunately, the labels applied are not always as benign as 'cat' and 'no cat.' Especially when discussing policing and surveillance, these labels can be challenging to apply in an objective, fair manner. This difficulty has significant ethical implications in a national security context. If these labels are applied inappropriately, then the algorithm will result in inappropriate future classifications (i.e., 'garbage in garbage out'). A website called ImageNet Roulette "is meant in part to demonstrate how various kinds of politics propagate through technical systems, often without the creators of those systems even being aware of them."[13] When you upload a picture to ImageNet Roulette, the algorithm detects faces and labels them. The labels are often racist, misogynist, and offensive. Young black males are identified as rapists and offenders, while white males are labeled managers and diplomats. What is disturbing is that this website used training data from the dataset which powers many of today's object recognition algorithms. Such applications in a national security context raise serious social, political, and ethical concerns.

1.3.2.1 Biases of the Past

Some of the biased data out there that can impact mass surveillance are historical criminal justice data. It is no secret that racism and racial profiling were, and still are, driving daily decisions by officers of many police

[13] https://imagenet-roulette.paglen.com/

departments. The data that we have regarding criminals, therefore, will be infected with these biases. For example, in criminal sentencing, it has been shown that "inmates with more Afrocentric features received harsher sentences than those with less Afrocentric features" (Blair et al. 2004). Were one to simply use the data from the past for an algorithm designed to make decisions regarding the sentencing of criminals, then the algorithms would simply reproduce the same bias. Concerning counter-terrorism, it is conceivable that bias might result from the fact that most terrorists and their associates in the past were Muslims of middle-eastern appearance (because extremist jihadists) but that in the present (as now appears to be the case in the US, at least) most terrorists and their associates are Anglo-Saxon whites in appearance (because of right-wing extremists).

Indeed, this kind of bias has happened. A report by ProPublica based upon the sentencing of 7000 people arrested in Broward County, Florida, US showed that black defendants were "77 percent more likely to be pegged as at higher risk of committing a future violent crime and 45 percent more likely to be predicted to commit a future crime of any kind" (Angwin and Larson 2016). This bias leads black defendants to receive harsher and longer sentences. This is real-world harm caused by the underappreciated role of training data in ML algorithms.

Data will also be needed to train ML algorithms used for mass surveillance. Detecting criminal activity will rely on what has previously been labeled as criminal activity. What data is chosen and how it gets labeled has a high risk of reinforcing biases of the past – resulting in unjust outcomes for certain groups of people.

1.3.2.2 Disparate Impact

The above is one form of what is called 'disparate impact.' Disparate impact means that although something results in higher performance overall, the benefits from such

performance are unequally distributed amongst different groups. It can also be that the adverse side effects disproportionately affect specific groups. For example, let's say an algorithm is used to stop people at the airport for extra scrutiny by the authorities. After using the algorithm for some time, it is found that 90% of those selected by the algorithm for additional scrutiny are black. This means that due to the algorithm, black people will be more heavily burdened with extra scrutiny than white people. With ML, past biases in the data are just one of several causes that can result in disparate impact.

In some cases, there probably is too little data from a specific group – resulting in the algorithm working poorly for that group. Facial recognition, for example, has been shown to misclassify darker-skinned females the most (up to 34.7 percent). For comparison, the same study showed that lighter-skinned males had a misclassification rate of 0.8 percent (Buolamwini and Gebru 2018). Similar issues with facial recognition have been established for younger and darker-skinned people in general (Klare et al. 2012). This would likely be so for members of terrorist groups, at least in the US and Europe.

When the Transportation Security Administration (TSA) claims that using facial recognition will speed up security lines at airports, it may indeed be true that overall, the lines are moving faster. However, this benefit may not be felt by those whose facial recognition technology does not work well on (dark-skinned and female). These left out people will also face closer scrutiny by immigration officials. Facial recognition is just the most discussed and studied form of ML being used for surveillance that could result in a disparate impact. It will be essential to make efforts to prevent this, and, most importantly, to check that algorithms being used are not resulting in such effects.

1.4 The Effectiveness of ML for Mass Surveillance

Any justification for the use of ML algorithms in mass surveillance will include something about the effectiveness of these algorithms. More specifically, it is a necessary condition for justification that these algorithms be effective. If they are faster and cheaper but are entirely ineffective, there isn't much to discuss. If facial recognition equipped smart CCTV cameras consistently deliver false positives and fail to provide true-positives, then we don't need to discuss privacy to make a case against their use, although they do constitute privacy infringements.[14] The case against their use can be made on efficacy alone.

This makes efficacy extremely important when it comes to any technology used in morally salient contexts like mass surveillance. ML comes with issues surrounding its effectiveness. Any development and use of ML for mass surveillance will have to take such matters into account in order not to waste a lot of time and effort. Furthermore, the use of an ineffective algorithm for mass surveillance means that there is no counterbalance for its harms.

1.4.1 Not Enough Data

Inefficacy in ML can be caused when ML is used to detect something for which there is very little data. For example, there are many proposals to use ML for detecting terrorists.

The amount of terrorists in the world is relatively small. Using the high end of the range given by a report by the US state department, we can estimate the total number of terrorists in the world to be about 100,000. That is 0.000014% of the world's population (US State Department

[14] For more on false-positives and false-negatives see Henschke (2019)

2018). Detecting terrorists within a specific group is even more difficult if we are forced to break this out by type of terrorist group (e.g., Jihadist vs. Far right). For machine learning algorithms, this is important for the training data set's balance, which will, in turn, have implications for the recall and precision of the resulting trained algorithm (see next section).

Trying to train ML algorithms to identify terrorists based on collected data would require that we have many examples that are associated with known terrorists. While there will indeed be some examples, the training data set will be highly imbalanced – meaning that the amount of non-terrorist related data far exceeds the amount of terrorist-related data. The algorithm could achieve a 99% accuracy rate by merely classifying all data as not terrorist-related (because so few data will be associated with terrorism). Imbalanced datasets are a subject of serious ongoing research, but the problem is far from solved (Johnson and Khoshgoftaar 2019).

1.4.2 Precision vs. Recall

The precision of an algorithm is measured by the total number of true-positives divided by the total number of true + false positives. If a smart CCTV camera were used to detect theft, for example, the precision rate would indicate the percentage of algorithm classified 'theft' incidents that were incidents of theft. A low precision rate means that many incidents labeled as theft were not theft. The recall rate of an algorithm indicates the total number of true positives divided by the total number of true positives + false negatives. The recall rate would show the percentage of actual thefts detected. A low recall rate means that many incidents of theft were not detected.

As can be seen from the description above, both recall and precision are desirable. However, they will always be in tension with one another. If you design and train the

algorithm to maximize recall, then you will have an algorithm that is good at correctly classifying all of the target cases but will most likely also many false positives. For example, if you wrote an algorithm to try and classify social media posts as hate speech or not and you wanted to maximize recall, then you would like to train the algorithm with a wide variety of hate speech examples. In your quest to capture every hate speech post, you are likely to also capture testimonials from people who have encountered hate speech, newspaper articles, etc. – leading to outcries of censorship. If, however, you want to maximize precision, then this leaves room for people posting hate speech to think of innovative ways to escape the algorithm. Who would have thought that the phrase "big luau" would be associated with far white extremists about killing police officers?[15] An algorithm capable of capturing this coded hate speech will also capture innocent Hawaiians notifying their friends and family on Facebook of their upcoming pig roast.

Think about an algorithm trained to Social media platforms have long had to walk this tightrope. As they improve the recall rate (e.g., they reduce the number of posts by terrorist groups), the precision rate declines (e.g., the number of seemingly related but entirely innocent posts wrongly taken down rises). Journalists writing about terrorism find their posts taken down and their accounts suspended. However, when they turn up the precision to prevent taking down innocent posts, they inevitably miss taking down horrific posts by terrorists.

1.4.3 Past vs. Future

Criminals, and especially members of sophisticated terrorist groups such as Al Qaeda, are good at changing the modus operandi (MOs) and, more generally, their habits once they realize that the state is using its knowledge

[15] See the article on the so called "Boogaloo Boys" by the Economist (Sweet 2020)

of those MOs and habits to catch them. When they figure out that cell phones are being monitored, they use 'burner' phones (pre-paid cell phones that are thrown away after a certain amount of use). Terrorists have taken to replacing and trading sim cards to frustrate the ability to monitor cell phone activity.

This cat and mouse game makes the use of ML incredibly tricky, as ML is most effective at detecting fixed targets. This is why many headlines are describing the success of ML at games (see, e.g., Thompson 2019). Games have a clear result in the end – win or lose (maybe draw) – and the rules of the game never change. ML is trained on past data. But in a national security context, the targets are not fixed. Cell-phone use patterns may give away criminal activity – but due to society changing the way communication occurs, the patterns of the future will change. For example, many now use messaging apps like WhatsApp instead of text messaging. This drastically changes the patterns of cellphone use amongst not only the general public but of criminals as well.

Furthermore, due to security concerns, many criminals, notably terrorists, have further changed their behavior by using security-focused applications like Signal (Detrixhe 2018). An ML algorithm trained on old data will be useless. Now there are specific techniques for evading ML-based detection. For example, it was shown that wearing a sign on your body makes one algorithm not classify you as a person (Vincent 2019). You can also put dots on your face and wearing specific makeup (Thomas 2019).

An over-reliance on ML can create a situation where criminals avoid detection, while innocents are constantly surveilled. This is not only ineffective but also unethical – as innocents are having their privacy violated without benefit to the state.

1.5 Meaningful Human Control over ML for Mass Surveillance

When using ML algorithms in morally salient contexts, there is great concern over how to keep human beings accountable and responsible for the algorithm's decisions. For a human to be held accountable and responsible, they must have had some sort of meaningful control over the situation. The concept of meaningful human control gained in importance in the literature regarding autonomous weapons systems (see, e.g., "Article 36" 2015, 36). It is now of concern whenever machines are deployed in morally salient contexts (contexts which could result in harm – broadly construed) (Robbins 2020; Santoni de Sio and van den Hoven 2018).

For mass surveillance, we want to make sure that people are not being labeled 'criminals,' 'terrorists,' 'suspicious,' etc. by an algorithm without meaningful human control. This is because being labeled one of those things could have harmful consequences, for example, being placed on a no-fly list. At the very least, it will result in increased surveillance of such a person, and this more targeted surveillance is caused by a process we do not understand. ML algorithms are opaque to their reasoning – that is, it is not clear why a particular decision was reached. Without ML, someone would have to give justification for more intrusive surveillance in the form of features or actions taken by the individual warranting such intrusive surveillance. With ML, one would only be able to point to the result of the algorithm. The issues of bias and efficacy highlighted earlier make such a situation unacceptable as ML may merely be using race, gender, or some other inappropriate correlation to deliver its output. To overcome this problem, there are a few proposals for ensuring that humans have meaningful human control over ML. Here we will highlight the significant proposals and their issues.

1.5.1 Veto Power

The most straightforward and most naïve proposal for establishing meaningful human control over ML is to give a human 'veto' power. That is, to put a human 'on the loop' to allow them to stop or 'veto' the machine's decision. In the Surveillance space, this could be a smart CCTV camera deciding that person A has committed theft. A human being could see that the decision had been made and could stop the process that starts as a result of that decision (e.g., authorities stopping and searching the individual). Another version of this (sometimes called human in the loop) requires that the human being consent to the decision made by the algorithm before the process continues.

At first glance, this keeps the human in control. However, as we take a closer look, veto-power offers the appearance of human control without anything 'meaningful' about it. When an ML algorithm classifies a person as a terrorist, for example, there could be a person in the loop or on the loop who can overrule that decision. However, on what basis would such an override be made? The algorithm had massive amounts of data at its disposal. Even if there is no evidence confirming the decision by the algorithm, could a human being reasonably say that the person was not a terrorist? ML is powerful in part because it will detect patterns and make associations that are impossible for a human to make. The default option is to confirm the algorithm's output. Without any information, then the human will simply confirm the decision without any evidence to support their decision.

Furthermore, the person placed in or on the loop will be set up to fall victim to human cognitive biases (Henschke 2019). The first being automation bias – in which humans fail to look for contradictory information when computers give solutions. When machines are widely used in a specific context to automate decisions, human beings who

previously made those decisions lose things like situational awareness, which are necessary to have to make these decisions. This dramatically reduces their ability to exercise their veto power (Cummings 2012). This problem is also associated with confirmation bias – the tendency to look for information that will confirm something rather than contradict it.

1.5.2 Explainability

Another option to establish meaningful human control is to force algorithms to provide explanations for their 'decisions.' The opacity regarding the factors that contributed to a decision by ML makes it difficult to judge whether or not that decision is acceptable or not (Robbins and Henschke 2017). There are many research projects devoted to achieving explainable AI (XAI) (see, e.g., Adadi and Berrada 2018). The idea is that with the decision by the algorithm, there would be an accompanying explanation which humans could use to understand better why the algorithm resulted in a particular output. This could better enable someone in or on the loop to determine whether or not that decision was acceptable.

For example, if an algorithm decided that someone was a terrorist and listed the factors to include the height of the person and the color of their skin, a human could determine that this was an unacceptable justification for placing someone on a terrorist watch list and veto the decision. While interesting, XAI falls short in establishing meaningful human control. First, explainable AI is a theoretical idea for which there is no current solution – and is therefore not a real option so far.[16] Second, if we already know the factors which should contribute to morally salient decisions, then it seems we should merely be automating these decisions the old fashioned way (Robbins 2019). Finally, the power of ML is

[16] Although promising progress has been made. See, for example, Wachter et. Al. (2017)

driven by its opacity. The considerations used by ML to make a decision are incomprehensible to human beings – that is why they can do things we cannot. Restricting its power to only look at human articulable factors would significantly reduce its capability (Robbins 2020).

1.5.3 Moral Machines

One idea gaining traction is to simply delegate moral responsibility to the machines themselves – making human control unnecessary. Research groups around the world have received significant funding to teach autonomous weapons systems ethics (Evans 2019). The hope is that ethical machines will not violate liberal democratic values – in this case, values of proportionality, necessity, discrimination, etc. Many others have been working on such endeavors for a long time (see, e.g., Wallach 2007).

There are a host of ethical issues that this idea raises, as well as grave concerns over efficacy. If the whole point of machines is to help to achieve our goals, and it is conceptually problematic to delegate moral responsibility or agency to them (Bryson 2010a). Machines cannot accept such responsibility even if we try to give it to them— they cannot be punished, nor do they 'care' about anything. Their mistakes will be made without feelings necessary for moral responsibility (Johnson 2006b). Giving ethics to machines further complicates what we were trying to solve in the first place: gaining control over these machines. Ethics adds a complex layer to an already complex machine – making them less predictable and, therefore, more difficult to control (van Wynsberghe and Robbins 2019).

The most devastating critique is that ethics is not something for which we have a solution. There is widespread disagreement about the list of, and interpretation of, ethical values. Ethics is not like chess, where we all agree about what checkmate is. Moral judgments will always be subject to disagreement. The

methodology for incorporating ethics into machines, then, will itself involving choosing between competing moral options about which reasonable people could disagree. Van Wynsberghe and Robbins argue that "no critical or unique operational function appears to be gained through the endowment of ethical reasoning capabilities into robots." Therefore we should "simply not do it" (van Wynsberghe and Robbins 2019).

1.6 Conclusion

This chapter has highlighted the many ethical and efficacy pitfalls that await those who wish to use ML for national security applications like surveillance. A product powered by ML, when delivered to a state institution, may already have serious baggage due to the training of the algorithm used. The data used to train it could be contaminated with bias, come from a problematic source, and be labeled in ways that diminish its efficacy, and cause certain groups to be disproportionately impacted.

In the context of surveillance, in particular, there are problems with efficacy since there may simply not be enough data to make accurate predictions. Practices surrounding criminal activity, and terrorist activity, in particular, change over time – causing ML trained on past data to be useless. Also, the precision/recall value hierarchy cannot be solved. Decisions about this in a surveillance context will always be subject to critical judgment.

In light of these issues and the moral salience of the context, human control over these ML algorithms is of the utmost importance. However, the solutions on offer so far are woefully lacking. Veto power offers the veneer of control without anything meaningful. Explainable AI has yet to be developed and may significantly reduce ML's abilities. Moral machines face many ethical issues of

their own and simply exacerbate the problem of human control.

The problems highlighted above may all seem to be insurmountable for the use of ML for surveillance. However, by acknowledging these possible issues, one can focus on decisions that significantly reduce the possibility of these issues. A police department in Massachusetts, for example, uses speech recognition to allow its officers to write reports and take notes on patrol. This enables them to surveille the area instead of having to have their head down writing. Officers claim that this helps prevent police ambushes and helps increase their ability to do their job (Condon 2018). While this solution is not as exciting as those offering to predict crimes, including terrorist attacks, or successfully generate profiles of terrorists and other criminals, it is incredibly beneficial. It does not fall victim to the many issues highlighted in this chapter. ML is no silver bullet for law enforcement or national security in general or counter-terrorism in particular. Treating it like a silver bullet will lead the state into the many pitfalls highlighted in this chapter.

Chapter 2.

National Security Intelligence and Bulk Data Collection Ethics[17]

2.1 Introduction

The rise of internet communications has necessitated an increase in digital national security intelligence collection (including counter-terrorism intelligence and military intelligence) – currently at a scale never seen before in liberal democracies. The Snowden revelations of 2013 revealed digital intelligence collection that was pervasive and perhaps illegal (Greenwald 2013a; Greenwald and MacAskill 2013). People around the world were shocked at the capabilities of the NSA to monitor their actions online. It is now 2020, and the intelligence collection practices revealed by Snowden have not slowed down. On the

[17] A version of this is to be published as: Robbins, S. "Bulk Data Collection Ethics" (forthcoming). In Miller, S. (ed), *Counter Terrorism: The Ethical Issues*. Edward Elgar. UK.

contrary, many of these practices are being enshrined in the law (Pieters 2016; Travis 2016; West 2018). Whether or not these practices are legal, it is essential to understand whether or not they are ethical – or how these practices can be conducted ethically. This involves identifying what makes these practices different from those that came before. Then one must highlight how this changes the ethical analysis.

Two broad ethical paradigms constrain the practice of intelligence. First, there is what is acceptable for law enforcement – which generally takes a case by case approach to evaluating the acceptability of collecting intelligence or surveilling a subject or subjects. Important considerations for law enforcement are: that there is reasonable suspicion or probable cause that the suspect (or suspects) are going to commit a serious crime, that the intrusion of their privacy is proportionate to the violations to citizens which will be the victims of that crime, and that the intelligence collection is necessary (i.e., there is no less intrusive alternative). Second, there is what is acceptable for national intelligence agencies to do during war – which in terms of intelligence collection means there are few constraints on their intelligence collection and analysis activities. Bulk data collection (BDC) for counter-terrorism purposes poses problems for each of these paradigms for two reasons. First, terrorism can and should be dealt with as a crime by law enforcement (Miller 2008); however, terrorist groups often target the state as a whole and, therefore, may require wartime tactics in response. Second, BDC, by definition, sweeps up large amounts of data on innocent people, which is not something typically allowed by law enforcement. This has created a murky situation concerning counter-terrorism intelligence collection and analysis. This chapter cannot solve this complex problem or, instead, set of issues; however, this chapter does provide some clarity on, and justification

for, constraints that ought to be imposed on one specific form of intelligence collection: BDC.

Contemporary scholars have frequently discussed the ethics of intelligence within a Just War Theory framework – those principles deemed necessary for the ethical initiation, conduct, and termination of war. Principles such as just cause, right intention, proportionality, last resort, etc. are now being used to evaluate the ethics of intelligence practices. These scholars are aware that war and intelligence practices are not the same kinds of activity (e.g., war is kinetic) and have made efforts to modify Just War Theory into a Just Intelligence Theory which accounts for the differences (Bellaby 2012; 2016; Gendron 2005; Macnish 2014; Omand and Phythian 2013; Quinlan 2007). My focus in this chapter is to apply some of the latest work in Just Intelligence Theory to BDC. This serves two purposes: first, to come to an understanding of the critical ethical issues surrounding the practice of BDC for intelligence purposes and, second, to highlight how Just Intelligence principles can be used to evaluate a specific intelligence program. There has been some formal use of Just Intelligence Theory to evaluate some intelligence programs, notably the use of torture to extract information. Moreover, there has been a relatively large public and journalistic outcry against BDC as an alleged violation of privacy rights. However, thus far, in the academic literature, there has been no comprehensive ethical review of the practice of BDC for intelligence purposes.[18]

2.2 Bulk Data Collection

To collect in bulk roughly means that the scope of collection will likely pick up many records that are not associated with current targets (Anderson 2016; Council

[18] Bellaby (2016) does give an in depth ethical evaluation of cyber-intelligence (broadly construed) with a couple paragraphs on what he calls "*en masse* collection" which is what he calls BDC.

2015). For example, the intelligence community (IC) may want all the records related to the current so-called Islamic State (IS) leader Abu Ibrahim al-Hashimi al-Qurashi. If the IC were only to collect records associated with him, then the IC is not collecting in bulk; instead, they are conducting a targeted collection. However, if the IC wants all records coming into and out of Syria because they think many terrorists are operating there, then the IC is collecting in bulk. There are many Syrians whose data will be collected who are not engaged in terrorist acts and who do not even interact with terrorists. This is significant from an ethical standpoint because the IC is knowingly collecting data on innocent people and doing so on a large scale. This will be important for evaluating BDC in terms of the ethical principle of proportionality (see below).

BDC is done in two different ways:

1. Bulk Interception: the practice of intercepting internet communications data which is in transit
2. Bulk Acquisition: the practice of acquiring bulk data from telecommunications and internet companies.[19]

Bulk interception is accomplished by placing fiber optic splitters on a telecommunications entry points. These fiber optic splitters copy the data and pass it along to intelligence agency infrastructure. These data will be filtered – to ensure that the data collection meets legal requirements[20] and also that as little irrelevant data ends up on agency servers as possible.

[19] Bulk interception and bulk acquisition are terms used by David Anderson in his review of the UK's proposed Bulk Powers Act which later became the Investigatory Powers Act (Anderson 2016).
[20] In the United States, for example, there must be minimization procedures to ensure that as little US person data as possible ends up on intelligence agency servers. See e.g. Blum (2008).

Bulk acquisition works in two ways. First, intelligence agencies can simply ask (or force) third party institutions to turn over data in bulk (i.e., data resulting from the application of some filter). Second, intelligence agencies may have back-door access to third party institution servers. The Snowden revelations revealed that such back-door access was given to the NSA by Google, Facebook, and others (Greenwald and MacAskill 2013).

In this chapter, BDC is also taken to be *prima facie* wrong, given it involves infringing the privacy rights of innocent citizens on a large scale. The purpose of this chapter is to understand what the conditions would have to be for its use to be justified.

Privacy or other rights of any given *targeted* person, i.e., a person who is an object of prior reasonable suspicion, cannot be the sole focus in the ethical evaluation of BDC. By definition, BDC is not targeted in this sense. Rather, it is the members of an entire group of people whose data will be collected to isolate members of that group for scrutiny. These groups are the result of filters being applied to the data passing through the internet. The filters themselves, then, are where the focus should lie for an ethical evaluation of BDC. It is these filters which delimit the set of potential "targets." Few would have objected to a filter that selects all data related to Osama Bin Laden. In the case of bulk collection, the filters are, by definition, much broader. These filters are what should be evaluated – in other words, the focus of this chapter is on understanding what might make the use of a particular filter morally justified and what might not do so.

2.3 Just Intelligence

As already mentioned, a prominent theoretical perspective in the field of intelligence ethics advocates adapting

Just War Theory (JWT) to evaluate intelligence collection and analysis. The primary reason for basing an ethics of intelligence on an ethics of war is that the conduct of both war and intelligence collection involves actions that are prima facie unethical. In war, you are killing people, destroying bridges and cities, holding people captive, etc. All of these things are ethically bad. However, there are cases when such actions are necessary, proportionate, and, more generally, morally justified. A country being invaded by another country should be able to defend itself – including shooting at their invaders. Just War Theory outlines principles that are held to be necessary and sufficient to justify going to war (*jus ad bellum*) and to justify the conduct of that war once it is being waged (*jus in bello*). Michael Quinlan argues that the practice of intelligence must also be justified and limited. In other words, there should be conditions that justify starting an intelligence program and limitations on how to conduct that intelligence program justly. Quinlan names these *jus ad intelligentiam* and *jus in intelligentia* (Quinlan 2007).

The reason for using a theory based on just war theory for intelligence collection is that intelligence collection involves harm and rights infringements that needs further justification. Intelligence collection can include listening in on private conversations, torture, deception, interception of communications, etc. All of these actions would also be ethically disallowed under normal circumstances.

Harms from BDC can be divided into two types: privacy infringements and restrictions on autonomy. The data swept up by an intelligence agency belongs to someone. An individual owns the information which that data reveals (Bellaby 2012). Prima facie, no one should be allowed to take this data. Of course, if this person is a known terrorist, then a state would be justified in collecting

any and all information about this person. The point is that a state needs to justify its actions concerning BDC because there is harm, or there are rights infringements associated with such intelligence programs. If the state fails to justify such infringements, then violations have occurred.

The autonomy of people – including citizens of the bulk data collecting state – can be restricted – intentionally or unintentionally – by BDC programs. Public knowledge of government BDC could affect the autonomy of innocent people whether or not their data is collected. The so-called "chilling effect" is when governmental regulation and policy not directed at certain activities deters individuals from carrying out protected activities (Robbins and Henschke 2017).

2.4 Just Bulk Data Collection

It is not the purpose of this chapter to justify the use of Just Intelligence Theory; rather, it is to use principles of Just Intelligence Theory to tease out ethical issues that arise due to the practice of BDC. In what follows, I use the JIT principles of just cause, proportionality, right intention, and proper authority to uncover issues that must be overcome to justify the use of BDC.

2.4.1 Just Cause

What would be a just cause for intelligence collection? As counter-terrorism is the most salient reason given in recent times for BDC, this analysis will be restricted to cases involving terrorism.[21] At first glance, it is clear that counter-terrorism is a just cause for an intelligence operation. If terrorists are indeed attempting to conduct attacks on citizens of a country, then that country has just cause to collect intelligence that would prevent

[21] However, this analysis will apply to any context where national security is at stake.

those attacks. Arguably, things might not be so simple for the reason that "the general threat of terrorism, the so-called War on Terror, for example, is too indistinct to offer any specific just cause for an operation" (Bellaby p. 313).

This gives us a generality problem for the just cause principle. 'Generality' refers to the problem that the justness of the cause changes depending on the scope of your view. As intelligence collection is primarily done to prevent threats from being realized, how specific that threat is characterized is essential. For example, if the IC has credible intelligence from an ISIS fighter captured in Syria that Bob Jones is going to attack Nebraska, then the IC has a particular threat that justifies the targeted collection of communications sent or received by Bob Jones. This is an obvious just cause for the intelligence operation. At the opposite end of the spectrum, if the IC knows that throughout their state's history, there have always been threats to national security, then they can inductively reason that there will be threats in the future. Therefore, someone might claim that the IC has just cause to collect intelligence on everyone in the world to prevent those unknown future threats from being realized. Since the IC doesn't know where the threats could come from in the future, no restriction on data collection would occur. This argument is spurious even if one is working with a reasonably broad definition of national security.

However, this is not a complete picture for two reasons. First, BDC occurs on a spectrum. At the most targeted end of the spectrum, BDC might consist of collecting all data from a small town known to be the home of terrorists. At the least targeted end of the spectrum, BDC might consist of collecting all data that, at some point, was physically in a location outside of the United States. The justification for a particular instance of BDC will depend

upon where it falls on this spectrum. Second, there is a conceptual issue regarding the point at which intelligence has been collected. On one account (further explained below), it seems as if the NSA, for example, gathers most of the data traveling through the internet – as most of the world's non-Chinese internet traffic flows through the United States. On the NSA's account of collection, the NSA collects a tiny fraction of the data traveling through the non-Chinese internet. The result of this analysis will affect when the just cause principle can be applied.

Regarding BDC occurring across a spectrum, two examples show the opposite ends of the spectrum. If the NSA were to place taps on the cables serving as the backbone of the internet (which they do) and collect ALL of the data streaming through them – thereby effectively collecting most of the communications data on the internet, then BDC is occurring at its most general. However, this (to the best of our knowledge) does not happen in practice. This is because of practical, legal, and (hopefully) ethical concerns. Practically, it would require a tremendous amount of storage and computing power to sift through all of that data. Of course, this practical issue arises with the data they do collect, but at a significantly smaller scale than if they collected ALL data on the internet. Legally, many countries have requirements that force intelligence agencies to ensure they are not collecting data on their own citizens (without a warrant). In practice, this means that filters are run on the incoming data to ensure that only the data legally allowed will be collected.

Now comes the conceptual issue of what counts as 'collection' as it is not simple in the case of BDC. When can data be said to have been 'collected' by an intelligence agency? It may be helpful to take a rudimentary look at an email which ends up in the hands of an intelligence analyst through BDC:

When the email is sent, it gets routed to the backbone of the internet run by (mostly) US communications companies (like, for example, AT&T). The communications company acts as the post office in that it makes sure the communication is directed towards the intended recipient. It is here at this first stage of the process (stage 1) that, for example, the NSA has a splitter on the fiber optic cables to copy the data. At this stage of the process, the data would have to be stored until filters could be run on it. At the next stage of the process (stage 2), the filters analyze the data and discard the data that does not match any of the filters. At stage 3, the data that makes it through the filters ends up on NSA servers for storage. Finally, at stage 4, an analyst queries the data resulting in the email (along with other data perhaps) being returned to the analyst who reads it.

With stage 1 above, it is clear that for some time, the email is stored on a government server – despite it being temporary. NSA owned equipment has possession of the data; however, at least as I have described the process, there is no potential for analysts to access that data. This is difficult as there is no analog for this in the physical world. Airport security can come close. When you put your bag on the conveyer belt, it now sits on airport security property. If the machine which selects baggage for inspection were automated (with no human in control), then this would be much like the situation with BDC. All bags must pass through, but only a few are passed on for further inspection. We would hardly say that our bags have been collected (or that our privacy has been infringed) simply because they are on the conveyor belt. But once that bag is directed away from all the other bags towards the inspection team, the bag has been 'collected.' In this analogy, the bag going through the machine is like the data in temporary storage – it rests on the property of the collectors. Still, it is inaccessible to them (again, provided that the baggage machine is automated).

Stage 1 is further complicated if analysts have access to the data stored temporarily. That is, analysts have access to the data before the filters have been applied. The Snowden revelations seemed to suggest that this was the case with a program called XKeyScore (Greenwald 2013b). It has been claimed that this temporary storage of all internet data lasted up to five days and effectively 'slowed down the internet' to allow analysts to query unfiltered data during that time. If this is the case, then the data has been collected. This results in a clear violation of just cause based on the fact that at this point there is no filter on the data. If there is no filter on the data, then the only justification available would be that there is a significant terrorist threat directed at the collecting state coming from the entire population in the world using the internet. Even for the United States, this is nowhere near true. So if XKeyScore exists as described by, for example, The Intercept (Lee et al. 2015; Marquis-Boire et al. 2015), then it fails to meet Just Cause.

The intervention at stage 2 appears trivial at first glance. It is merely the state of the data as filters are being run on it. It should look like a series of questions: Did this data come from Syria? No. Iraq? No. Is it encrypted using tools known to be used by terrorists? No. etc. If any of the questions results in a yes, then the data moves onto long term or permanent storage. I include stage (2) in my discussion because I want to highlight the difference between using these filters and running complex pattern matching algorithms. Filters appear to be simply automating a human process. If one were to print out all of the emails passing through the internet, a human could check to see which of the emails matched one of the filters. Computers speed this process up, but a human being could quickly double-check each communication if need be. This is opposed to complicated computer algorithms that attempt to find patterns in the data or

make predictions on that data. For example, a deep learning algorithm could be trained on all of the communications associated with terrorism (previously), and use that to classify future communications in terms of their connection with terrorist communications or other terrorist actions. A deep learning algorithm basically forms heuristics based on the training data to form judgments about new cases. This is no longer the automation of a human process; rather, it is a novel process that could well be opaque to human minds. What can be said about algorithms like these being run on the data in temporary storage? Earlier I argued for evaluating the filter for just cause, but in this case, the filter is opaque to evaluation. The computer scientist who created the original algorithm would not even be able to explain how the algorithm classified a particular communication as being associated with terrorism.

What is just cause supposed to apply to if the filter is opaque? In this case, it seems clear that just cause is violated as all data is being searched for suspicious (or terrorist) patterns. In the case where these filters are applied, the discarded data is not analyzed in any reasonable sense of the word. These filters are simply not articulable – meaning that they cannot be described in human language. However, an argument could be made that if the algorithm is better at classifying communications in terms of their connection to terrorism than the articulable filters are, then the fact that they are not articulable should not be a reason not to use them. In other words, using machine learning algorithms could be better for privacy because they are more accurate in their classifications. In the healthcare context, for example, machine learning algorithms are much more accurate in classifying moles on skin as cancerous than dermatologists. The fact that those algorithms are opaque about how they classify these moles should not stop those

algorithms from being used (Esteva et al. 2017; Presse 2018; Robbins 2019).

Is the classification of communications in terms of their connection to terrorist activity like the classification of possibly cancerous moles? To this, I can definitively say 'no.' First, the reason that the IC is collecting data in bulk is in part because of the changing communication tactics of terrorist groups. In the case of moles, which makes them cancerous does not change. In theory, the algorithm will only get better over time (and this has so far proven to be the case for moles). However, the classification of communications into those relevant to terrorist activity, and those not relevant will change drastically over time. How the Irish Red Army (IRA) communicated is very different from how ISIS communicates. ISIS has changed how they communicate over time as technology has changed (and with the knowledge that they are being surveilled). What this means is that in the case of moles, an algorithm trained with data on thousands of moles and then put to work classifying new moles will be fixed with regard to efficacy. The nature of cancerous moles is not going to change, so the target is the same every time. With respect to potentially terrorism-relevant communications, there is a moving target. This is so for three reasons: first, technology is continuously progressing, which changes the way we as a society communicate; second, terrorist groups of the future may communicate drastically differently than terrorist groups of the past; and third, terrorist groups know they are being surveilled and actively modify the way they communicate to thwart intelligence agencies.

All of this means that it will be incredibly difficult to say that an algorithm is better at classifying communications than an articulable filter. This, added to the fact that the criteria used by the algorithm are opaque, makes it unethical to use such complex algorithms

as filters for BDC. Without evidence about the efficacy of the algorithm, the IC would lack the ethical justification necessary to use it as the algorithm will cause harm or rights infringements. To make matters worse, these harms and rights infringements occur without human accountability. That is, no human being can be called upon to justify any data collection caused by the algorithm. Of course, the humans responsible for deciding to use the algorithm at all can still be held accountable; however, who would accept such responsibility when it is not possible to determine the efficacy, or the underlying logic, of the algorithm? I do concede, however, that if an algorithm that showed itself to be robust and effective in its classification of internet communications as being associated with terrorism, then using such an algorithm in place of an articulable filter would be acceptable. For the reasons stated above, this situation is highly improbable.

At stage (3), it is common to classify the data as collected. In this case, the data rests on government servers with access given to analysts under institutional constraints. This data is justifiably collected when there is evidence that there is a terrorist threat being organized or planned by the group described in the filter resulting in the collected data AND that this threat is directed at the state collecting that data. While this may satisfy just cause, whether or not it is proportionate to the threat is another question (see below).

2.4.2 Proportionality

Is BDC a proportionate response to the threat of terrorism? Talk of proportionality with respect to going to war (*jus ad bellum)* is stated as a condition that "the destructiveness of war must not be out of proportion to the relevant good the war will do." (Hurka 2005, 35). The principle of proportionality is also used for evaluating the just conduct of war (*jus in bello),* albeit in the

context of the principle of discrimination and the principle of military necessity. The principle of discrimination states that the action (e.g., bombing an arms factory) must not target non-combatants (however this is cached out) as a means or as an end. This prevents states from intentionally killing an enemy's civilians to demoralize them. According to the principle of necessity, the action must serve a military purpose. According to the *jus in bello* proportionality principle, the (unintended) deaths of innocent civilians, while permissible if militarily necessary, must not be disproportionate in the sense that the number of innocent deaths is disproportionate relative to the importance of the military objective (Hurka 2005).

Applying these principles to intelligence collection, and BDC, in particular, is difficult. For one, in war, intelligence collection is often used to determine if people are combatants or non-combatants – which would help in establishing that an action is proportionate. What intelligence is the IC to use to show whether intelligence collection is being directed at non-combatants (or innocents or civilians)? Kevin Macnish states the difficulty by saying: "Surveillance is often carried out in order to determine innocence or guilt, and so the status of the surveilled prior to the act of surveillance is frequently unknown." (Macnish 2014, 151).

Looking specifically at the intelligence program of BDC, one can quickly see that the evaluation of proportionality hinges on empirical data. The extent of the harm done by BDC is challenging to determine before it has been carried out – the same goes for the extent of the good it will achieve. How pervasive is the so-called chilling effect described in section 2.3 above?

The extent of these kinds of chilling effects is relatively difficult to quantify. How are we to know whether or not someone has changed their internet behavior

– much less what the reason for that change was? A Pew research center poll, however, did conclude that 25% of Americans have changed their online behavior due to perceived government surveillance (Gao 2015b). Depending on the methods used, the harms could be even more widespread – and more difficult to quantify. The bulk acquisition of data from third-party institutions – especially when it pertains to back-door access and data retention – could result in the diminished trust in participating institutions. Edward Snowden, in an interview with the New Yorker, explicitly told people not to use Dropbox, Google, or Facebook because of their susceptibility to intelligence collection (The New Yorker 2014). This, in turn, could harm the profits of third party institutions and the US economy itself. This can be shown in advertisements where companies brag about where their servers are located because they are located in countries without close intelligence ties to, for example, the US. ProtonMail (an encrypted email provider) states that:

> *As ProtonMail is outside of US and EU jurisdiction, only a court order from the Cantonal Court of Geneva or the Swiss Federal Supreme Court can compel us to release the extremely limited user information we have.*

It will be necessary going forward to understand the harms to third party institutions as a result of BDC. Harms like these must be taken into account in any calculation of proportionality. These harms would then have to be weighed against the efficacy of the program – or the good that it will do, which of course, is another empirical matter.

Moving onto the principle of discrimination – according to which an intelligence program should not target innocents (or non-combatants) as a means or an end. *BDC* IS the intentional collection of data from innocents. Intelligence agencies know that most of the data collected are from people not associated with terrorism in any way.

One could argue that these innocents are being used as a means for collecting data on terrorists – in which the principle of discrimination would be violated. On the flip side, one could argue that the only intention with respect to BDC is to collect data on terrorists – all of the data associated with innocents is incidentally collected, a.k.a. collateral damage. While the amount of people included in the collateral damage is relatively large, it may still be proportionate depending on how one characterizes the level of harm.

2.4.3 Right Intention

Limiting the focus of the chapter to BDC for counter-terrorism should make right intention a no-brainer. If the government intends to prevent terrorism, then right intention should be of little concern. Much like Just Cause above, the situation is not so simple. There are cases of just cause being met while failing at right intention. In war, at least, it is an easy task to find examples. Think of the bombs dropped on Hiroshima and Nagasaki. If (and that is a big if) just cause was met because of the threat and aggression of Japan if the intention of the United States was to deter? the USSR, then the United States did not meet Right Intention.

The same situation can happen with BDC. There may be a clear threat in Afghanistan of terrorism directed at the United States – a threat that constitutes just cause for BDC. However, the intention of the collecting state may be to glean information helpful to influence elections there. If that were the case, then the collection state does not meet right intention.

What complicates right intention, however, is when and how often it should be applied. In the case of dropping bombs on Hiroshima and Nagasaki, it is quite evident when right intention should be applied – when the decision is made to drop the bombs. Analogously right intention should be applied to the decision to create a filter that results

in BDC. However, there is a time dimension which complicates this in two ways: (1) the filter will continue to collect data long after the decision was made to use that filter, and (2) the collected data will be stored long after that decision.

To illustrate the problem with (1) above, let us act as if BDC was a tactic being used to combat the IRA, and the British intelligence agencies had just cause to collect all of the data coming into and out of Ireland. The IRA is no longer the threat it once was, so not only would British Intelligence have to re-evaluate just cause, but they may have a problem with right intention as the British intelligence agencies may leave the filter because the data could be useful in the future.

2.4.4 Proper Authority

In Just War Theory, it is often stated that the proper authority to initiate war resides with a state. For intelligence and surveillance, this is no longer the case. In certain circumstances, reporters, private investigators, individuals, corporations, etc. all might have the proper authority to conduct an intelligence program. For example, reporters might gather intelligence on the comings and goings of lobbyists visiting a government office to report on the influence of those lobbyists. However, as this evaluation of BDC is restricted to national security, such nuances can be ignored. A reporter should not be allowed to collect bulk communications data from the internet to get a story.

One could simply go along with traditional Just War Theory and claim that the only proper authority for BDC is the state. If this is the case, then a problem arises because, in practice, there are many third-party institutions collecting data in bulk. The practice of bulk acquisition is about the state copying data which has already been collected by third party institutions – either by request or by back-door access. The question becomes one of

whether or not the third party is then collecting bulk data as part of an intelligence collection and analysis program.

In many instances, this is not the case at all. Telecommunications and internet companies store a lot of data that is necessary to conduct their business. Google doesn't store your email on their servers for national security. They store your email so that you have access to it. If they deleted your emails, then they would no longer be an email service provider. Setting aside the bulk nature of the collection, it can be seen that intelligence agencies requesting data from third parties is benign (provided it meets other just intelligence principles). There is nothing inherently wrong with obtaining data from third parties. If Osama Bin Laden had a Gmail account, it would, and should, be expected that the NSA ask Google for those records – and it would, and should, be expected that Google provide them.

Things get more interesting if we look at forced data retention policies – in which laws mandate that third party institutions retain data they may not typically retain for counter-terrorism (or national security). Now, the third party institution is engaging in BDC as an intelligence program. This fails the principle of proper authority. Not only this, but now all of the data which has been retained that usually would not be should be included in our evaluations of just cause, right intention, and proportionality.

Once the IC is forcing third parties to retain data for the explicit purpose of something like counter-terrorism, then what is the difference with the government collecting all of that data themselves? One difference will be who has the keys to the retained data. If the government does not have back-door access, then they must rely upon the third party institution to hand over that data. Unless they have the freedom to refuse, then the difference here

is one of procedure rather than outcome. In the case of back-door access, then to an analyst, the procedure and the outcome could be the same.[22]

This problem is exacerbated when it is understood what the broad purpose of retaining such data would be. The purpose is, purportedly, national security. So the government faces a dilemma concerning the value of this data. Either the data is essential for national security, or it is not. If the data is necessary, then the storage of that data should not be contracted out to third party institutions. This is both because of the security risk of third parties being hacked and the blurring of institutional aims that such storage causes. Blurring these institutional responsibilities could damage the company's reputation as well as make it easier for those wishing to evade detection to choose other institutions. If the data is not essential, then they should not be forcing third party institutions to retain such data.

The reason that the IC should not be contracting the collection and storage of data for purposes of national security to third party institutions is those third party institutions are directed by commercial ends and have motives other than the protection of national security. If, for example, Google were to simply stop collecting specific data points because it helped their public image (and therefore advertising profit) while still retaining the data they did collect, they would do it. They may be pressured to do that since they are primarily (though not exclusively) focused on maximizing returns for their shareholders. This could happen despite those data points being essential to making the retained data useful for national security purposes. If this is possible, then the

[22] Of course the procedure could be different in that law or policy require different levels of justification for access to third party institutions data.

government should not be contracting such retention of data out to third party institutions.[23]

2.5 Conclusion

This chapter has used the latest work in Just Intelligence Theory to evaluate the practice of BDC in liberal democracies for intelligence purposes. Using Just Intelligence Theory forced me to understand precisely what it was that would be the object of evaluation – the filters used to funnel data into government servers – as well as to tease out some important ethical issues surrounding the practice. Most importantly, this evaluation points us to some crucial constraints which should be placed on this practice. These constraints included: not using artificial intelligence as filters; not allowing consumer companies like Google and Facebook to act as intelligence agencies (collect data for the sole purpose of counter-terrorism); collected data must be tied to a filter and deleted when the justification for that filter no longer holds, and; data must only be used for the legitimate purpose originally provided for its collection.

This evaluation is just a start; however, it points to constraints that are not currently in place with regard to BDC. Furthermore, this chapter starts with the premise that BDC is a valuable tool in the fight against terrorism. This may not be the case. If this tool turns out to be ineffective, then it should not be used with or without the constraints outlined above. The point is that if intelligence agencies want this tool in their arsenal, they should be using it in a way that conforms to liberal democratic principles and values. Having a just cause and right intentions to collect data in bulk, which is proportional to the threat and conducted by a proper authority, would be a good start.

[23] For a general argument against contracting out intelligence operations to corporations see Roper (2010)

Chapter 3.

The Value of Transparency

Bulk Data and Authoritarianism[24]

3.1 Introduction: Disrupting Relations between Citizens and the State

Following the revelations by Edward Snowden about widespread state-sponsored surveillance programs (Greenwald 2014; Harding 2014), some fear that liberal democracies are at risk of descending into authoritarianism. "Obviously, the United States is not now a police state. But given the extent of this invasion of people's privacy, we do have the full electronic and legislative infrastructure of such a state…These powers are extremely dangerous" (Ellsberg 2013). The revelation of comprehensive government surveillance programs implies that liberal democracies are about to become authoritarian police states. Here we need to stress the relationship

[24] A version of this was previously published as: Robbins, Scott, and Adam Henschke. 2017. "The Value of Transparency: Bulk Data and Authoritarianism." Surveillance & Society 15 (3/4): 582–89. https://doi.org/10.24908/ss.v15i3/4.6606

between individual privacy and individual freedom. While surveillance obviously infringes privacy it can also compromise freedom, given that privacy is in large part an aspect of freedom. After all, *control* of one's persons data, for instance, is an important component of individual freedom. Moreover, if surveillance by government is intrusive and on a large scale, then it can compromise the freedom of the citizenry as a whole, i.e. a power imbalance between government and citizenry can emerge and, thereby, the potential for creeping authoritarianism.

An essential driver of the erosion of individual freedoms and privacy has been the counter-terrorism policies of the US and other liberal democracies since 9/11. Since the 9/11 attack, the US has engaged in detention without trial of suspected terrorists at US facilities in Guantanamo Bay, bulk data collection of the personal data of US citizens, most of whom are known to be innocent of terrorism or, indeed, of any crime (revealed by Snowden) and, in more recent times, the introduction of controversial new technologies, notably face recognition technology. As has been pointed out by numerous commentators, there is a danger that counter-terrorism policies designed to protect liberal democracies will undermine them by compromising their core principles and values. In this chapter, our focus is on bulk data collection and analysis, in particular.

Despite many legitimate concerns about state overreach and worries about the national security organs in liberal democratic states, the US, UK, and other similar countries remain worlds apart from somewhere like North Korea: perhaps people like Daniel Ellsberg are worried over nothing? However, looking at modern surveillance technologies suggests that there is a disruption of relations between citizen and the state.

Snowden's revelations shed light on the notion that a fundamental shift had occurred between liberal democratic states and their citizens – In the US, for example, the National Security Agency (NSA) had access to vast amounts of information about its citizens, while the citizens knew nothing of this. "These bureaucratic ways of using our information have palpable effects on our lives because people use our dossiers to make important decisions about us to which we are not always privy" (Solove 2004, 9). This is highly important as the relationship between liberal democratic states and their citizens is central to them *being* liberal democracies – should such a state cease representing those citizens, then it can no longer properly be called a liberal democracy.

The information communication technologies (ICTs) that enable comprehensive surveillance disrupt the state/citizen relation due to their 'opacity'; what Snowden showed was a 'revelation' not so much because it publicized specifics about state surveillance programs, but because these programs were being done in the name of these state's citizens without their knowledge, albeit to protect them from terrorists and other malevolent actors. "While the government, via surveillance, knows more and more about what its citizens are doing, its citizens know less and less about what their government is doing... Democracy requires accountability and consent of the governed, which is only possible if citizens know what is done in their name" (Greenwald 2014, 208–9). We ask if a state can be representative of the will of its citizens if those citizens do not know what the state is doing.

'Representativeness' is core to liberal democracies. Here, "public authorities are bound by their own rules and can only exercise their powers in a lawful way. All powers must derive from the constitution… [implying] the important fact that the government is accountable and that its actions must be controllable, and thus transparent"

(Gutwirth and de Hert 2006, 64). This representativeness has elements of legitimacy and control in it. First, on the social contract view, the legitimacy of the government is dependent on the state *actually representing* the view of those it governs. Second, for our purposes here, the governed *need to know* what is being done in their name, such that they can remove legitimacy should it become apparent that the state is not representing the will of the governed (Altman and Wellman 2009, 3–6). The strong public condemnation of the US surveillance programs following Snowden is evidence that many US citizens did not know what was being done in their name. Bulk, open-ended surveillance programs are problematic specifically for this – the information gathered is potentially limitless in its use and who can use it.

This points to the more profound disruption being caused by ICTs. A recent article in the *Scientific American* asked whether democracy will survive big data and artificial intelligence (Helbing et al. 2017). The authors' concerns stem from the ignorance of consumers regarding how increasingly sophisticated ICTs can be used by states and private companies. What we are suggesting here is that ICTs are disrupting the relations between the citizen and the state by giving the state unparalleled access to information about its citizens. In contrast, the citizens are not comparably informed about what the state is doing or what it knows. This is not to say that there might not be benefits to citizens arising from the state's access to their personal information; as conservative politicians in the US are quick to point out, the US has not suffered any major terrorist attack on the homeland since 9/11. However, it is to say that there is what we can call an 'informational deficit'; the state's knowledge about its citizens substantially surpasses what the citizens know about the state. While there has always been some informational deficit between what a state does and what its citizens know, the worry here is that the new

technologies provide so much more information about its people, without a corresponding increase in the citizens knowing about the state.

Furthermore, these technologies are increasingly complex to the point of being opaque to citizens' understanding. Not only are the ICTs challenging to understand due to their complexity, but artificial intelligence (AI) used to process bulk data collected by the state can be inherently opaque – even to operators of the technology. The value of representation must be realized in the design of these ICT systems.

3.2 Designing For Representation: Ensuring and Assuring the Will of the Citizens

Liberal democracies are, by definition, not authoritarian. "The original impulse of the liberal tradition, found in Locke and Kant, is the idea of the moral sovereignty of each individual. It implies limitations on how the state can legitimately restrict the liberty of individuals even though it must be granted a monopoly of force in order to serve their collective interests and preserve the peace among them" (Nagel 2002, 63–64). In the liberal tradition, we forgo certain rights to have collective goods such as security.[25] Core aspects of liberal democracies protect against authoritarianism; they have processes to *ensure* and *assure* us that they are not authoritarian. On ensurance, a state cannot *become* authoritarian. And on assurance, the citizens *know* the state is not authoritarian. We suggest that the ICT systems that support bulk-data collection can be designed with processes that ensure and assure a state's citizens for the sake of representation. In short, much of what security agencies need in the way of bulk data collection to ensure that the keep the citizens safe from terrorists

[25] Here, we are agnostic about security as an intrinsic good, or an instrumental good that protects other collective goods.

and other malevolent actors can be permitted; however, this can and should only be permitted in a manner consistent with the social contract.

The methodology we use is a specific articulation of Value Sensitive Design (VSD). VSD starts with the idea "that aims at making moral values part of technological design, research, and development. It assumes that human values, norms, moral considerations can be imparted to the things we make and use" (van den Hoven 2007, 67). This is particularly relevant when looking at the relation between citizens and the state and how those relations can be disrupted by technologies. "If we want our information technology - and the use that is made of it – to be just, fair and safe, we must see to it that it inherits our good intentions. Moreover, it must be seen to have those properties, we must be able to demonstrate that they possess these morally desirable features" (van den Hoven 2007). What follows is a brief conceptual investigation (as per VSD) of the value of representation and end-norms to ensure and assure that liberal democracies do not become authoritarian.

One step in actively designing for a particular value is to specify that value – here, we are mainly concerned with the value of representation in liberal democracies. We suggest that a focus on two design elements can go some way to translating representation into the design requirements of state surveillance technologies. Here, we consider that surveillance (and other state-based bulk data collection programs) should be designed such that they ensure the will of the citizens is represented, and that the citizens are assured that the state's actions are representative of their will. This draws from the notion of a values hierarchy in VSD, in which the design of an ICT system brings in 'end-norms' for the sake of the ultimate value being designed for. "End-norms in design then may refer to properties, attributes or capabilities

that the designed artefact should possess" (van de Poel 2013a, 258).

The two end-norms that we consider of prime importance to bulk data and representation are to *ensure* and *assure* that the state represents the will of its citizens. 'Ensurance' is the attribute of a system in which the design features ensure that an end is either being achieved or not frustrated by technologies. To ensure that a surveillance program is neither being used in the citizen's name without their knowledge nor is being used against the state's citizens, the citizens or their representatives[26] must be informed about how bulk data is collected, how it is being used, and who has access to it and so on. Here, insofar as bulk data about citizens is collected and potentially used against a state's citizens in secret, then ensurance has not been met, and representation has been undermined. Reducing informational deficits can *ensure* that there is no slide into authoritarianism. That is, if the citizens disagree with what's being done in their name, then they can withdraw support for the state. This capacity to withdraw support is core to representative democracies. However, in cases of informational deficits, the citizens don't know what's being done, so they can't know when to withdraw support.

The second end-norm is concerned with assuring a state's citizens that the state is representing their will. Here, the process is about the confidence that citizens have that their will is represented. A "[g]overnment therefore has to explain through the media the rationale for the strategy it is following and convey a sense of where and why it is balancing the benefits from additional security with all the costs of providing it" (Omand 2010, 18). An informational deficit between the citizen and the state

[26] Here, 'representative' refers to something like the US Foreign Intelligence Surveillance Act (FISA) courts in which oversight is achieved through those acting on behalf of the citizens.

is concerning because such deficits can cause changes in citizen's behavior. A Pew Research Center poll showed that a quarter of Americans have changed their behavior with regard to technology due to US government surveillance (Gao 2015a). This is the so-called "chilling effect" – when governmental regulation and policy not directed at certain activities deters individuals from carrying out protected activities. Currently, bulk data collection in the name of national security is directed at terrorists and malevolent state actors. However, if being concerned about government surveillance, a citizen is deterred from participating in legitimate political organization and activism, then the chilling effect has taken place. The Pew Research Centre report shows this "chilling" has affected the behavior of a quarter of Americans. What is interesting is that this effect takes place even if policies to protect citizen's privacy exist. This is because citizens don't have the knowledge needed to feel assured: on this, good policy must not only exist but be known by the citizen to exist. Informational deficits can keep the chilling effect in place.

3.3 The Instrumental Value of Transparency

Transparency is the solution to informational deficits. However, transparency is not a value in itself; rather, it is instrumental for realizing other values such as representation that are important. Transparency can be "an ethically "enabling" or "impairing" factor (Turilli and Floridi 2009). Thinking about transparency in this way prevents radical approaches in which the only option is full disclosure – which would undermine the values (e.g., safety and security), which the government is set up to realize. The goal for governments is to use transparency to enable ethical values that are lacking due to the informational deficits outlined above.

Transparency can be used to *ensure* that privacy is not being overridden without justification and authorization

and to *assure* citizens that there are appropriate policies in place. However, the simple fact that this information is available does not mean that assurance has been realized. The process of effectively disclosing this process is essential:

> Information transparency should disclose not only information but also details about how such information has been produced. Such details are a necessary condition for verifying the consistency between the ethical principles endorsed at the time of producing information and the ethical principles that information transparency should enable *(Turilli and Floridi 2009, 109)*.

For transparency to fulfill its instrumental value, information needs to be disclosed to the public in a way that can realize the value of an assured citizenry. For instance, the processes by which this information is chosen, prepared, and redacted may also need to be disclosed. Before we can articulate how to realize representation through transparency in the design of ICTs, which enable bulk data collection, we must understand how these ICTs reduce transparency.

3.4 Bulk Data Collection and Opacity

Bulk data collection has made the quest for transparency a moving target. This is so for two reasons: first, the technology (both infrastructure and code) behind bulk data collection is difficult for the public to understand; and second, some algorithms used to process bulk data are intrinsically opaque due to properties of contemporary approaches to AI and machine learning.[27] Both of these reasons make the regulation as well as the ensurance and insurance discussed earlier challenging to realize.

3.4.1 Technical Opacity

While debating net neutrality – the idea that internet service providers should treat all content equally –

[27] For more discussion on opacity and machines see (Burrell 2016)

Senator Ted Stevens famously said that the internet is "a series of tubes" (Belson 2006). While there is some debate about how accurate that metaphor is, it represents an opacity that has severe consequences for the regulation of the internet. Senator Ted Stevens, despite his technical illiteracy, gets a vote on important legislation – and people who are equally technically illiterate vote him into office.

With respect to bulk data collection, for example, much has been discussed about the NSA tapping the backbone of the internet (Greenwald 2014; Kravets 2013). The NSA has partnered up with telecommunications companies and the intelligence agencies of other liberal democracies (the Five Eyes[28]), to place fiber optic cable splitters on the cables that serve as the backbone of the internet. This effectively puts a "tap" on the internet. The information which flows through this tap must be filtered[29] and stored on government servers. To properly understand what this means to citizens, one would have to understand how much of the internet's traffic flows through these taps, what filters are in place to ensure data is not collected from citizens, and how much 'incidental' citizen data is collected. A further point which is important but will not be explored here is what institutional and legal arrangements are made to prevent governments from bypassing restrictions on collecting citizen data by merely obtaining the data from other countries.[30]

[28] 'Five Eyes' is an intelligence sharing agreement between the US, UK, Canada, New Zealand, and Australia

[29] By US Law there must be a procedure in place for minimizing US person data. The FISA court approves these procedures once a year.

[30] The FISA Courts in the US, for example, are particularly attentive to this citizen/non-citizen (or, more precisely a US person/non-US person) distinction. Which points to a third form of opacity, legal opacity. However, we do not have space to cover legal opacity in this paper.

While there have been speculation and educated guesses about how much data flows through these taps[31], the filters used and how successful they are at both preventing citizen data collection and preventing terrorist attacks is unknown to the public. Without an understanding of how much of their privacy is being "incidentally" invaded. No matter how effective this technological solution is with regard to terrorism, there is no way that the public can be *assured* that the state is not overreaching. Furthermore, citizens cannot consent to such a program without having some degree of knowledge about how effective it is. For example, if it were true that these technologies would have prevented 9/11 (a hypothetical that is widely disputed)[32], then citizens may conclude that it is worth it to have this program even if it has the potential to invade their privacy.

3.4.2 Algorithmic Opacity

Artificial Intelligence (AI) uses machine learning algorithms, which are excellent at categorizing things. For example, Google image search can categorize images quite accurately.[33] And now there are proposals coming both from academia and private companies for how to use AI to combat terrorism (Aviv and Aviv 2009; Frenkel 2017), discover illegal immigrants, catch tax evaders (Hemberg et al. 2016), etc. Governments are feeding algorithms bulk collected data so that algorithms can make decisions about us.

AI methods are being increasingly employed to enhance a system's ability to reach decisions about large data sets. AlphaGo, developed by Google, uses a method called deep

[31] http://sniffmap.telcomap.org/ tries to show how much data the NSA and its partners intercept.
[32] Former FBI chief Robert Mueller claims that bulk data collection would have prevented 9/11 (Roberts 2013) while CNN national security analyst and journalist Peter Bergen forcefully argues against this idea (Bergen 2013). An in depth read about this can be found in a 2015 New Yorker article (Schwartz 2015)
[33] There have been some embarrassing mistakes however. See (*BBC News* 2015)

learning to "learn" how to play the game Go. When AlphaGo makes a move, not even the programmers understand why it made the move that it did. This is algorithmic opacity – opacity which is a result of the properties of an algorithm itself. This makes the decisions made by AI algorithms opaque to even those who are technically literate.

3.5 Restoring Representation: Transparency by Design

Having established that transparency plays a vital role in preventing a government's slide into authoritarianism by instrumentally supporting ensurance and assurance, governments must do what they can to limit technical and algorithmic opacity. "For this, the state would have to provide an appropriate regulatory framework, which ensures that technologies are designed and used in ways that are compatible with democracy… Individuals would then be able to decide who can use their information, for what purpose and for how long" (Helbing et al. 2017).

What follows is a brief technical investigation (as per VSD) of the challenges and corresponding design requirements associated with the systems involved in government bulk collection of meta-data. This is not a detailed, exhaustive solution to the problem of opacity. It is an important step towards the ideal of a representative, transparent, and legitimate liberal democracy.

3.5.1 Technical Transparency

Making the technical aspects of bulk data collection more transparent to a technically illiterate public is important for assuring that governments are not overreaching. Non-Governmental Organizations in the US like the Electronic Frontier Foundation (EFF)[34] have gone some way to doing that by providing infographics and easy

[34] https://www.eff.org/

to read descriptions of how governments bulk collect data. However, the stance of the EFF is decidedly directed at problems with the government's bulk data collecting programs. The government should make a concerted effort to take the lead in explaining the what, when, how, and why to keep the public assured that there is no government overreach.

A design requirement which would go some way in realizing the value of transparency would be to audit bulk collected data and remove citizen's data. This solution could involve human auditors, or because there is so much data, algorithms. This would enable the ability to report on how much incidental data is collected and the processes in place to remove that data. While citizens would not necessarily understand the technical processes behind bulk collection, knowledge of the results of these processes would help tremendously.

This should include the efficacy of these processes. How good are these processes at preventing terrorism? The US government at least has been silent about this.[35] Without this knowledge, citizens cannot begin to balance the values of privacy and security. While making transparent the details of how exactly a specific terrorist plot was prevented may compromise security, general reporting on the success of bulk collection programs is necessary to realize transparency.

To sum up, two specific requirements result if the government were to design these technologies which realize the value of transparency – which is instrumental to both ensurance and assurance: first, the ICT must include the capability of auditing the system for incidentally collected data associated with citizens; and second, there

[35] In a very recent congressional hearing, the NSA attempted to give examples of the success of bulk data collection – none of which clearly showed that this data helped to prevent attacks in liberal democratic countries (Savage 2017)

must be the ability to report on its success in preventing terrorism. Institutional and legal arrangements should be made to distribute this information to the public (in a way that does not compromise the success of the ICT). Without these processes, the will of the public is not effectively represented.

3.5.2 Algorithmic Transparency

Making the decisions of AI algorithms transparent is a hot computer science topic.[36] Researchers and companies seem to understand that decisions made by algorithms will not be tolerated if we cannot understand them. No one wants to be prevented from getting on a plane because an algorithm put them on the terrorism No-Fly list without an explanation. The nature of some of these algorithms (e.g., deep learning) makes a solution to opacity extremely difficult, and we should not expect that this will be accomplished anytime soon.

The solution, therefore, is to use such algorithms for specific situations in which it is acceptable not to have an explanation or to supplement the decision of the algorithm with human oversight. Placing someone on the No-Fly list, for example, should not be solely decided based on an algorithm that can offer no explanation. A restriction of one's rights is a moral decision, and only a human being can accept the moral responsibility which comes along with such a decision.[37]

The design requirement which comes out of this will help realize the value of ensurance. Ensurance is realized in this situation if the government is prevented from using these kinds of algorithms for moral decision making. Policy should be written to show that this is the case, and this policy should be made public so that citizens are

[36] See for example a recent PEW Research Center report (Rainie and Anderson, Janna 2017)
[37] For more on this discussion see Bryson (2010b) and Johnson (2006a)

assured. Only with such assurance will representation be realized.

* * *

This brief technical investigation helps to move liberal democracies closer to realizing the instrumental value of transparency. Transparency is essential for ensurance that liberal democracies cannot become authoritarian. Transparency is also crucial for an assured public – a public confident that the property of ensurance has been met. VSD is key to responding to the disruptions caused by ICTs between liberal democratic states and their citizens. By specifying representation as a value and highlighting connections between transparency and technology, we can design ICTs for democracy.

Chapter 4.

Critiquing the Reasons for Making Artificial Moral Agents[38]

4.1 Introduction

Robots perform exceptionally well at clearly defined tasks like playing chess, assembling a car, classifying images, or vacuuming your floor. Increasingly, however, robots are being assigned more general tasks that require more than one skill. A sentry robot designed for perimeter protection, for example, is supposed to be "able to detect shapes and motions, and combined with computational technologies to analyze and differentiate enemy threats from friendly or innocuous objects–and shoot at the hostiles" (Anderson and Waxman 2012). Moreover, it is envisaged that combatants might be robots and, therefore,

[38] A version of this was previously published as Wynsberghe, Aimee van, and Scott Robbins. 2019. "Critiquing the Reasons for Making Artificial Moral Agents." Science and Engineering Ethics 25 (3): 719–35. https://doi.org/10.1007/s11948-018-0030-8

be able to distinguish enemy combatants from civilians or, in the case of counter-terrorism operations, terrorists (dressed as civilians) from innocent civilians; such robots would shoot dead enemy combatants and terrorists but not innocent civilians and, it is suggested, do so more reliably than human combatants (Arkin et al. 2012). For robots like these to execute their function, they require algorithms. These algorithms controlling robots are becoming increasingly autonomous and often require artificial intelligence (AI). As autonomy in robots and AI increases, so does the likelihood that they encounter morally salient situations. As of 2017, robots are and will continue to be designed, developed, and deployed in morally salient contexts; from robots in the hospital lifting or bathing patients to robots in the military assisting with bomb disposal, intelligence gathering, or even, as suggested above, killing enemy combatants and terrorists.

The Executive Summary of the International Federation of Robotics[39] shows a marked increase in robot sales across every sector from 1 year to the next, including a 25% increase in the total number of service robots sold in 2015 alone. These robots can be used to save lives, to assist in dangerous activities, and to enhance the proficiency of human workers. Many industry leaders and academics from the field of machine ethics—the study of endowing machines with ethical reasoning—would have us believe that robots in these and other morally charged contexts will inevitably demand that these machines be endowed with moral reasoning capabilities. Such robots are often referred to as artificial moral agents (AMAs). In this chapter, the variety of reasons offered by machine ethicists in favor of AMAs are challenged. We ask: are the

[39] For more on this see https://ifr.org/ifr-press-releases/news/world-robotics-report-2016.

given reasons adequate justification for the design and development of AMAs?

From the academic domain, a variety of scholars in the fields of ethics and technology and robot ethics have argued against the development of AMAs (Tonkens 2009; Bryson 2010a; Johnson and Miller 2008; Sharkey 2017). What is currently missing from the debate on AMAs is a closer look at the reasons offered (to society, academics, the media) by machine ethicists to justify the development of AMAs. This closer inspection is compulsory given the amount of funding allocated to the development of AMAs (from funders like Elon Musk) coupled with the number of attention researchers and industry leaders receive in the media for their efforts in this direction.[40] Moreover, the stakes are high because the resulting technology could create novel demands on society, questions about what counts as an AMA, whether they are deserving of citizenship,[41] and whether they are morally responsible for their behavior or not. In other words, a machine with moral reasoning capabilities might be thought to deserve moral consideration in the form of rights or protections (Coeckelbergh 2010; Darling 2012; Gunkel 2014).

To examine the justifications for AMAs, this chapter begins with a description of the field of machine ethics: what it is, the terminology used, and the response to machine ethics found in the literature by robot ethicists and scholars in the field of ethics and technology. In subsequent sections, the reasons offered in favor of developing robots with moral reasoning capabilities are evaluated. It is argued that each of the reasons lack both empirical and intuitive support. The burden of proof is

[40] For more on the popular news articles see: (Deng 2015; The Economist 2012; Rutkin 2014).
[41] Robot Sophia of Hanson Robotics, first robot granted citizenship in Saudi Arabia, see (Gershgorn 2017; Hatmaker 2017)

thereby shifted to machine ethicists to justify their pursuits.

4.2 Machine Ethics

Summarized by machine ethicist Susan Anderson, the "ultimate goal of machine ethics is to create autonomous ethical machines" (2007, 15). The term machine ethics was first used by Mitchell Waldrop in the AI Magazine article "A Question of Responsibility" (1987). In 2005 the AAAI held a symposium on machine ethics, which resulted in the edited volume *Machine Ethics* in 2011 by Susan Leigh and Michael Anderson (Anderson and Anderson 2011). The field may be referred to by other names, e.g., machine morality. Still, for this chapter, machine ethics is a field of study dedicated to the computational entity as a moral entity.[42]

There are several phrases and terms for discussing robots with moral reasoning capabilities (e.g., moral machines, implicit vs. explicit ethical agents).[43] For this article, however, the term artificial moral agent (AMA) will be used for consistency and clarity.[44] This restricts the discussion to robots capable of engaging in autonomous moral reasoning, that is, moral reasoning about a situation without the direct real-time input from a human user. How this might be done, and whether or not this can be achieved in practice, are questions that go beyond the scope of this chapter (these are the questions underpinning the field of machine ethics itself). Rather, the interest of this chapter is in targeting the reasons offered in support of developing such machines.

[42] For more readings on machine ethics see Wallach and Allen (2010), Anderson and Anderson (2007; 2011), Anderson (2011), Moor (2009; 2006), Scheutz (2016), and Allen et al. (2006).
[43] For more on this see Wallach and Allen (2010), Moor (2009; 2006).
[44] The concept and notion of artificial moral agents has built momentum as a thought experiment and/or a possible reality. For a rich and detailed discussion of AMAs we recommend the following: (Allen et al. 2005; 2000; Floridi and Sanders 2004; Himma 2009; Johnson and Miller 2008; Nagenborg, n.d.; Wiegel 2010)

What a robot or machine would act like if it were to think ethically is a central feature in the 1950 works of science fiction writer Isaac Asimov. Asimov, who coined the term 'robotics' (the study of robots), is best known for his work articulating and exploring the three laws of robotics (Asimov 1963). In short, these three laws were a kind of principled or deontological approach to embedding ethics into a machine. Through a series of short stories, Asimov reveals the difficulty and nuances of robots acting ethically because each ethical principle conflicts with another in many situations (e.g., lying to protect someone's life) to such a degree that experience, wisdom, and intuition are required to come to a solution or resolution of the conflict. His stories highlight the struggle to define ethics in a computational form.

From the academic domain, a variety of scholars in the fields of ethics and technology and robot ethics have argued against the development of AMAs. On the one hand, scholars insist that the technology ought to be designed in such a way that responsibility distribution remains "tethered to humans" (Johnson and Miller 2008). Similarly, computer scientist Joanna Bryson argues that robots ought to remain in the instrumental service of humans, as slaves if you will, meeting the needs of their human users and intentionally designed not to be a moral agent (Bryson 2010b). This claim is predicated on the assumption that humans will own robots and, as such, will be responsible for the consequences of the actions and outcomes of that robot. On the other hand, philosopher Ryan Tonkens argues that given the impossibility of finding universal agreement concerning the ethical theory used to program a machine, the initiative is moot (Tonkens 2009).

Outside of these arguments, robot ethicist Amanda Sharkey outlines the misappropriation of the use of 'ethical' in the quest to make moral machines and insists on the creation of "safe" machines instead. In the same line of

thinking, Miller et al. argue that responsible development requires careful use of terminology and representation in the media (Miller et al. 2017).

The above arguments are still waiting to be adequately answered by the machine ethics community. However, the purpose of this chapter is to question the positive reasons offered by the machine ethicists *for* building AMAs. These reasons have not yet been thoroughly evaluated, and a closer inspection of them reveals a lack of sufficient justification. Given the high stakes of the research and development in question coupled with the current speed of (and funding for) machine ethics initiatives, these must be addressed now.

4.3 Reasons for Developing Moral Machines

Machine ethicists have offered six reasons (found in the literature) in favor of the development of moral machines. These are not stand-alone reasons; rather, they are often intertwined. Part of the reason it sounds so convincing (at first glance) is because of their interdependency rather than the strength of any reason on its own. Disentangling these reasons shows their dubious foundation and allows one to challenge the endeavor of machine ethics.

4.3.1 Inevitability

> *Robots with moral decision making abilities will become a technological necessity (Wallach 2007).*
>
> *[Artificial Moral Agents] are necessary and, in a weak sense, inevitable (Allen and Wallach 2014).*

Machine ethicists claim that robots in morally salient contexts will not and cannot be avoided, i.e., their development is inevitable (Anderson and Anderson 2010; Moor 2006; Scheutz 2016; Wallach 2010).

First, what exactly is meant by morally salient contexts is unclear. For some researchers, this would include

contexts such as healthcare, elder care, childcare, sex, and, notably for the concerns in this work, the military, the police, and intelligence agencies, notably in counter-terrorism operations —where life and death decisions are being made on a daily (or hourly) basis (Arkin 2009; Lokhorst and van den Hoven 2011; Sharkey 2016; 2008; Sharkey and Sharkey 2011; van Wynsberghe 2012; Sharkey et al. 2017). There is no question that robots are entering these service sectors. The International Federation for Robotics Executive Summary of 2016 tells us that "the total number of professional service robots sold in 2015 rose considerably by 25% to 41,060 units up from 32,939 in 2014" and "service robots in defense applications accounted for 27% of the total number of service robots for professional use sold in 2015". Moreover, sales of medical robots increased by 7% from 2014 to 2015.[45]

For others, the morally salient context is much broader than a pre-defined space or institution:

any ordinary decision-making situation from daily life can be turned into a morally charged decision-making situation, where the artificial agent finds itself presented with a moral dilemma where any choice of action (or inaction) can potentially cause harm to other agents (Scheutz 2016, 516).

From the above quote, Scheutz is saying that a morally charged situation can arise at any moment if someone could be harmed through (in)action of a robot. This thin description of a morally charged decision-making situation adds further ambiguity to the discussion, namely (1) what level of autonomy does the robot have, and (2) what definition of harm is Scheutz talking about? There seems to be an assumption being made in the above quote concerning the robot that the robot must choose action or

[45] For more on this please refer to: https://ifr.org/downloads/press/02_2016/Executive_Summary_Service_Robots_2016.pdf.

inaction, and thus that the robot must be autonomous. According to Scheutz, then, any autonomous robot interacting with a human user that has the potential to harm its user should be endowed with moral reasoning capabilities. What would Scheutz have us do with industrial robots that possess divergent levels of autonomy, work with humans in their presence, and for which it has already been shown that the robots can bring severe harm or sometimes death to humans? Scheutz's position would imply that industrial robots as well ought to be developed into AMAs.

Consider also the definition of 'harm' that ought to be adopted. Is it only physical harm to the corporeal body and mind that is the object of discussion here, and if so, what about the robot's or AI algorithm's ability to collect, store and share information about its users in a home setting? Considering the real possibility that home robots will be connected to the Internet of Things (IoT), which holds the potential for hackers, companies, governments (foreign and domestic), and terrorist groups not related to the robotics company to access personal data from users for malevolent purposes. The harm that can come from the misappropriation of one's data has proven to be noteworthy of late: people can be refused mortgage loans, defrauded, stalked, blackmailed, harassed online, subjected to political propaganda, or worse, as in the case of those whose personal safety or even life depends on ensuring their identity and location is kept confidential, e.g., domestic abuse victims, police informants, and witnesses. If harm is to be extended to include the risk of one's digital information, and interaction with a machine that might cause harm demands that it be endowed with ethical reasoning capacities, then one must concede that every device that one interacts with in a day (your tv, phone, fridge, alarm clock, kettle, etc.) ought to have such capabilities. Thus, Scheutz's position leads to the conclusion that any technology that

one interacts with and for which there is a potential for harm (physical or otherwise) must be developed as an AMA, and this is simply untenable.

Second, a distinction must be made between *being in a morally charged situation*, on the one hand, and *being delegated a moral role* on the other. Consider animals used for therapeutic purposes in an elderly care facility; one would never demand that a dog placed in this context would need to reason ethically because of its role in therapy and the potential for harm in this context. Indeed the dog would be trained to ensure a degree of safety and reliability when interacting with it but would the dog be a *moral dog* in the end?[46]

With this thought in mind, let us say the discussion will be limited to an examination of a *morally salient context* to contexts such as the military and healthcare, which are often thought of as morally salient, and agree that it is inevitable that robots will be placed within these contexts. In this case, there is a different, more nuanced problem that can be put into the form of a dilemma: when placed in a morally salient context, either machines will be delegated a moral role or they will not. If one chooses the first horn—that the machine is delegated a moral role—then one must accept that it is inevitable that machines will be delegated a moral role in addition to the inevitability of the machine being in this morally salient context. However, this is simply not the case. There are plenty of machines operating in morally salient contexts that have not been delegated a moral role and are providing a valuable service. Consider, for example:

[46] See also the work of van Wynsberghe illustrating how robots in healthcare need not be delegated roles for which ethical reasoning and/or moral responsibility are required. (2012; 2016a; 2013; 2016b). Furthermore there are existing frameworks and applications for realizing ethical values in technological design. See e.g. (Friedman and Nissenbaum 1996; Nissenbaum 2001; van de Poel 2013b; van den Hoven 2007; van Wynsberghe and Robbins 2014)

An algorithm that uses satellite imagery to detect terrorist training camps. When the algorithm detects a new training camp, it raises an alert for a human operator to investigate. The human being is still in charge and retains all of the responsibility for decision making. If the flagged image, on closer inspection, is merely a wedding celebration, then the human can ignore the algorithm's warning rather than initiating the procedure for a drone strike. If one agrees with machine ethicists, then one should accept that it is inevitable that the moral role of choosing to ignore the initial classification or initiating a drone strike – in this example reserved for the human - will be assigned to the machine. While this is probably unnecessary and most likely harmful, the point is that there is simply no reason to believe that this is *inevitable* – even though this algorithm operates in a morally salient context.

If, however, one takes the other horn of the dilemma, then the claim is as follows: robots will inevitably be in morally salient contexts without being delegated a morally salient role. The problem with this is that there is little that is new here. Microwaves and coffee machines exist in the counter-terrorism field offices with no need for moral reasoning capabilities; this horn should be of little interest to machine ethicists. In short, there is not any evidence to suggest that it is inevitable that there will be a need for machines with moral reasoning capabilities regardless of whether or not they function in a morally salient context.

4.3.2 Artificial Moral Machines to Prevent Harm to Humans

For many scholars, the development of moral machines is aimed at preventing a robot from hurting human beings. To ensure that humans can overcome the potential for physical harm, a technological solution is presented; namely, to develop AMAs:

> *the only way to minimize human harm is to build morally competent robots that can detect and resolve morally charged situations in human-like ways (Scheutz 2016).*

The line of reasoning here is pretty straight forward in that: "it is clear that machines…will be capable of causing harm to human beings" (Anderson and Anderson 2010), and this can be mitigated by endowing the robot with ethical reasoning capabilities. This also speaks to the interconnection of the reasons in favor of AMAs; robots are inevitable, robots could harm us, therefore robots should be made into AMAs.

It is unclear that AMAs are the solution to this problem, however. There are plenty of technologies capable of harming human beings (e.g., lawnmowers, automatic doors, curling irons, blenders); the solution has always been either to design them with safety features or to limit the contexts in which a technology can be used. An elevator door has a sensor so that it does not close on people, lawnmowers have a guard to protect us from their blades, and ovens have lights to warn us when our stovetop is hot. One does not usually use barbeques indoors or chainsaws in daycare centers. Machine ethicists are the first to suggest endowing technology with moral reasoning capabilities as a solution to problems of safety.

Furthermore, machine ethicists may agree that their true pursuit is for safe robots. Then, of course, there is no reason to use the word 'moral.' Notions such as values, rights, freedoms, good vs. bad, right vs. wrong, are central to the study of morality and ethics and form the basis for a discussion of competing conceptions of the good life. One may believe that the values of safety and security are fundamental to achieving the good life; however, ethics cannot be reduced to these issues. So if AMAs are simply a solution to possibly harmful machines, then *safety*—not *moral agency*—is the object of debate.

In this case, the word 'moral' is a linguistic 'trojan horse'—a word

> *that smuggles in a rich interconnected web of human concepts that are not part of a computer system or how it operates (Sharkey 2012, 793)*

The concept of *moral* machines or artificial *moral* agents invites or more strongly requests that the user believe the robot may care for him/her or that the robot can experience feelings. For robot developers, this could increase the desirability of the robot and, therefore, profits. However, this is problematic for the public in that it invites a kind of fictive, asymmetric, deceptive relationship between human and robot.

The LS3 robotic pack mule used by the US military, for example, must operate safely. Like an elevator that stops short of crushing something put in between its doors, the LS3 should stop or move around an object in front of it. This, however, is nothing new. Technologies are built with many features designed to protect people from harm. Calling what the LS3 robot does when it goes around a US military serviceperson rather than running them over 'ethical reasoning' is misrepresenting what is happening.

Thus, machine ethicists must either distinguish what makes their machines "moral" above and beyond "safe," or they must stop using the word "moral" as the word is not appropriate—only the most reductionist account of morality would equate it with preventing harm.

4.3.3 Complexity

> *as systems get more sophisticated and their ability to function autonomously in different contexts and environments expands, it will become more important for them to have 'ethical subroutines' of their own (Allen et al. 2006, 14)*

The idea behind using complexity as an argument in favor of AMAs is that robots are and will increasingly become so complex in terms of their programming that it is no

longer possible to know what they will do in novel situations. This uncertainty results in the impossibility of the engineer to predict every scenario, and as such, it will not be possible for the engineer to predict the robot's actions. The hope is that AMAs will ensure that the outputs of these complex machines will not be morally bad ones. Consequently, one cannot foresee a morally problematic situation and pre-program what the robot should do. Instead, authors who use the complexity argument to promote the development of AMAs claim the robot needs to have moral competence to govern its unpredictable actions in the inevitably unpredictable and unstructured human environments that the robot will be placed.

First, using complexity as a reason for developing AMAs assumes both that there will be complex robots and that such robots ought to be placed in contexts for which this complexity (i.e., unpredictability) could cause problems for human beings.

Next, of importance for this issue is the context within which the robot will be placed. In other words, this problem can be mitigated simply by restricting the context within which these machines are used. For example, designers of Google's complex machine AlphaGo may not have any idea what their machine will do next (which move it will make in the notoriously difficult game of GO); however, this is not an ethical or moral problem because the context (the game of GO) is restricted. Its complexity will not pose a problem for us.

One may argue that human beings are unpredictable and can cause harm to other human beings. The solution has not been to prevent the delegation of moral roles to human beings. One might ask: why treat machines differently? While it is outside of the scope of this chapter to engage in a debate on just how predictable humans are, it can be noted that concerning serious moral values—killing, non-

consensual sex, harming innocent people for fun, or, in the case of terrorists, for a real or imagined just cause—society places restrictions on unpredictable human beings (i.e., imprisonment). Humans may be unpredictable in terms of what they will do next, but most of us assume that a random person will not intentionally cause us harm.

4.3.4 Public Trust

Other machine ethicists argue that making AMAs will increase public trust: "Constructing artificial moral agents serves at least two purposes: one, better understanding of moral reasoning, and, two, increasing our trust and confidence in creating autonomous agents acting on our behalf" (Wiegel 2006). There has been talk in the media expressing concerns surrounding AI and robotics—voiced by the likes of Elon Musk and Steven Hawking (Cellan-Jones 2014; Markoff 2015). Rather than preventing the development of robots that are the source of these fears, "machine ethics may offer a viable, more realistic solution" (Anderson and Anderson 2007).

This line of thinking assumes that if robots are given moral competence, then this will put the public at ease and lead to public acceptance. It should be noted here that acceptance differs from acceptability. As an example, the public may *accept* geotagging and tracking algorithms on their smartphone devices, but this does not mean that such privacy breaching technologies or the lack of transparency about their existence are *acceptable* practices for upholding societal values.

Some essential clarifications are needed when discussing trust as a concept. Traditionally speaking, trust is described as an interaction between persons or between a person and an institution, and so on. For scholar John Hardwig, trust can be placed in people, processes, and knowledge (Hardwig 1991). In more recent years, scholars are discussing a new form of trust; trust in algorithms (Simon 2010). This new form of trust is most commonly

referred to as 'algorithmic authority' and is described as a practice of placing confidence in the decisions made by an algorithm (Shirky 2009). Wikipedia is an example of this form of trust as it requires trust not in persons but in the algorithms regulating the content on the website.[47]

If trust is broken, the result will be feelings of disappointment on the part of the truster. These resulting negative feelings are what relates trust to the concept of reliability: if either one is misplaced, the result is oftentimes feelings of disappointment (Simon 2010). Trust is distinguished from reliability in the intensity of the emotions experienced afterward; "trust differs from reliance because if we are let down we feel betrayed and not just disappointed" (Baier 1986; Simon 2010). Relatedly, Simon claims that one cannot speak of trust for socio-technical systems but rather of reliance: "we usually do not ascribe intentionality to unanimated objects, which is why we do not feel betrayed by them" (p 347). Hence, we do not trust unanimated objects; we rely on them.

With the formulation of Hardwig in mind—trust can be placed in people, processes, and in knowledge. When it concerns placing trust in robots, one must ask: who, or what, are machine ethicists asking the public to trust: the algorithm directing the robot, the designer; or, the development process?

If the public is being asked to trust the algorithm, then one must consider that:

> *unfortunately, we often trust[48] algorithms blindly. Algorithms are hidden within a system. In most cases we are not aware of how they work and we cannot assess*

[47] This form of trust may also be referred to as procedural trust (Simon) as it concerns trust in the process through which knowledge is created rather than in actions of persons.
[48] The word trust is used here because it comes from a quotation; however, it should be noted that the authors are inclined to use the phrase 'rely on' instead.

their impact on the information we receive. In other words: algorithms are black-boxed (Simon 2010).

Consequently, if the public is being asked to trust an algorithm and it is considered a black-box, then, as Simon rightly asserts, it must be *opened*—the way it works, the decisions made in its development, and alternatives—must be made transparent and subject to scrutiny.

If, however, the public is being asked to trust the designer, then designers and developers ought to develop an enforceable code of conduct (perhaps in the form of soft law). Again, transparency of this is required for the public to have the knowledge required for trust.

Last, if the public is being asked to trust the process through which the robot is being developed, a kind of procedural trust, then standards and certifications must be developed to once again provide the user with the knowledge required to place trust in the process through which the robot was developed. Examples of such procedural trust are FairTrade, ISO, GMOs, and so on.

In any case, it is important to point out the inconsistency between the promotion of AMAs for reasons of complexity and reasons of trust: it is inconsistent to expect unpredictability in a machine and to expect trust in a machine at the same time. While this may not be the case for people—one might trust persons who are at the same time unpredictable—more clarity is needed in understanding who/what society is being asked to trust and what level of (un)predictability one can assume.

4.3.5 Preventing Immoral Use

In the 2012 American science fiction comedy-drama movie "Robot & Frank," there is a compelling story of how a retired cat burglar convinces his robot to help him enter the business once again. The story raises the question about human-robot interaction not in the sense of safe or reliable interactions but rather should the robot be

capable of evaluating a human's request for action. Thus, another reason put forward for the development of AMAs can be stated as: preventing humans from misusing, or inappropriately using, a robot requires that the robot be developed as a moral machine and can thus prevent misuse of itself.

The main problem with this reason has to do with the potential to constrain the autonomy of humans. It's not always clear what is the right thing to do, and frequently, context is required for this (Miller et al. 2017). Consider, for example, a case when police witness a man with a suspicious package on a train in London. It is thought that this person is a suicide bomber. The police direct their autonomous robot policeperson to shoot the man in the head to prevent the detonation of the bomb – thereby saving many lives. However, the robot is programmed, to prevent its misuse, to not shoot into crowds because of the danger of killing bystanders. Should the robot policeperson be programmed in this deontological manner? Meaning that no matter the consequences, it should never shoot into a crowd. Or should the utilitarian principle apply here – meaning that the robot would be allowed to shoot into the crowd considering the gravity of the situation? While this example might steer us in favor of the utilitarian principle, much hinges on how grave the situation would need to be for it to be considered ethical to shoot into the crowd. Most importantly, should the robot be delegated the task of making that determination?

Consider another example where misuse is unclear. If an older adult at home wants to have a fourth glass of wine and asks his/her robot to fetch it. If the robot fetches the wine, is the robot being misused in so far as it is contributing to poor health choices of the user? Or is the robot 'good' in so far as it fulfilled the request of its user. Presenting scenarios like these is meant to show the

difficulty in determining the right or the good thing to do. And yet if one is claiming that robots should be involved in the decision making procedure, it must be evident how a 'good' robot is distinguished from a 'bad' one.

4.3.6 Morality: Better with Moral Machines

Endowing the robot with the capability to override or edit a human's decisions draws us into the discussion of the robot as a superior moral reasoner to a human. Computer science Professor James Gips suggested back in 1994 that "not many human beings live their lives flawlessly as moral saints. But a robot could" (Gips 1994, 250). Also along the same lines, Professor of Philosophy Eric Dietrich has suggested that:

> *humans are genetically hardwired to be immoral…let us – the humans – exit the stage, leaving behind a planet populated with machines who, although not perfect angels, will nevertheless be a vast improvement over us (Dietrich 2001)*

The assumption here is that a robot could be better at moral decision making than a human, given that it would be impartial, unemotional, consistent, and rational every time it made a decision. Thus, no decisions would be based on bias or emotions; no decision would be the result of an affinity towards one person (or group of people) over another. More importantly, the robot would never tire but would have the energy to be consistent in decision making: to make the same choice time after time.

This line of reasoning to promote AMAs is also often invoked when speaking of robots in military contexts. In particular, computer scientist/roboticist Ronald Arkin discusses the power of autonomous military robots for overcoming the shortcomings of humans on the battlefield (Arkin 2009). These robots would not rape or pillage the villages taken over during wartime and would be programmed as ethical agents according to the Laws of Just war or the

Rules of Engagement. Professor Arkin states the goal of his AMA project explicitly: "creating a class of robots that not only conform to International Law but outperform human soldiers in their ethical capacity" (Arkin 2008).

There are some general concerns with this reason. First, the underlying programming which will enable machines to reason morally implies that one has an understanding of moral epistemology such that one can program machines to "learn" the correct moral truths—or at least know enough to have AMAs learn something that works. This gets complicated as there is no moral epistemology that does not have serious philosophical objections and therefore presents a barrier to being reduced to a programming language.

Machines could only be better if there is some standard of moral truth with which to judge. This implies that there are objective moral truths in a moral realist sense and further that it is possible to know what they are. This is opposed to error theory (the idea that there are no moral truths at all—so nothing to know), and moral skepticism (there are moral truths, but it is not possible that we, as humans, can know them).

Furthermore, based on the above quotes, it seems that the moral truths that machines would be better at knowing are truths that are independent of human attitudes. Russ Shafer-Landau calls these stance-independent moral truths (Shafer-Landau 1994). If—and that is a big if—there are stance independent moral truths whereby the truths have no dependence upon human desires, beliefs, needs, etc. then there are objections to how one could come to know such truths (Finlay 2007). If a machine were built, which did somehow discover moral truths that have previously yet to be discovered (because morality would be a lot easier if we simply knew the moral truths), then one would have to accept on faith that machines are better than we are.

The moral consistency promised by machine ethicists is only a public good if the moral truths are known in advance—the opposite of the situation human beings find themselves in. For, as shown in previous sections, AMAs are argued to be needed because one cannot predict the kind of situations or moral dilemmas they will face. But this is not a chess game where the outcome is a win or a loss. An autonomous car that drives off a cliff—killing its one passenger—in order to save five passengers in another car would not be a clear cut situation that everyone could agree was the correct decision. Indeed, books are written about that very decision and human disagreement about what should be done (i.e., the trolley problem) (see, e.g., Greene 2013).

Lastly, this all presumes that human emotions, human desires, and our evolutionary history are all getting in the way of our moral reasoning—causing it to be worse than it could be. Some include moral emotions as a necessary part of moral judgment and reasoning (Kristjánsson 2007; Pizarro 2000; Roeser 2010). If this is so, then AMAs would require emotions—something not even on the horizon of AI and robotics.

Let us say that there are moral principles and that humans can know what they are. So there is a standard with which to judge AMAs. Furthermore, let us also assume they live up to their promise and are better moral reasoners than humans. It might then make sense to outsource our moral decisions to machines. This would assume that being good at moral reasoning is not a necessary part of a human being's good life. Aristotle believed leading a moral life and gaining a moral understanding through practice was necessary to lead a good life (Aristotle et al. 1998). Many contemporary philosophers agree. Outsourcing our moral reasoning to machines could cause an undesirable moral deskilling in human beings (Vallor 2015). The point is that it is not clear at all if machines were better

moral reasoners than us that this would be a good reason to use them. Added to this, to make such an assumption is to assume we have an understanding of morality and the good life that we may not.

4.3.7 Better Understanding of Morality

Finally, machine ethicists sometimes argue that developing robots with moral reasoning capabilities will ultimately lead to a better understanding of how humans reason morally (not necessarily how they *should* reason):

> *the hope is that as we try to implement ethical systems on the computer we will learn much more about the knowledge and assumptions built into the ethical theories themselves. That as we build the artificial ethical reasoning systems we will learn how to behave more ethically ourselves (Gips 1994)*

In short, regardless of the resulting machine, the very process of attempting to create such a machine would benefit humans in so far as we would learn about ourselves and our moral attributes (Gips 1994; Moor 2006; Wiegel 2006).

The most critical consideration in response to this claim is that while various ethical theories may well inform human moral decision-making to a greater or lesser extent in different contexts, speaking generally, they are by no means the only factor in play. Therefore the work doesn't help understand *human* morality. Experiments in moral psychology show us that human morality is deeply influenced by irrelevant situational factors (Doris 1998; Merritt 2000), is driven by emotion (Haidt 2001; Haidt and Joseph 2008), and influenced by our evolutionary past (Street 2006). To be sure, there is an intense debate in the literature concerning each of these studies. The point is that human morality, in the descriptive sense, is dependent upon many complex factors, and building a machine that tries to emulate human morality perfectly

must use each of these factors combined rather than rely on ethical theory alone.

4.4 Conclusion

In this chapter, the reasons offered by machine ethicists promoting the development of moral machines are shown to fall short when one takes a closer look at the assumptions underpinning their claims. While autonomous robots and AI can and should be used in morally salient contexts, this need not require that the robot be endowed with ethical reasoning capabilities. Merely placing something in an ethical situation, like a heart monitor in an ICU hospital ward or a robot sentry in a military zone, does not also demand the thing to reflect on its course of action in terms of ethically salient features. The power of such robots in said contexts can still be harnessed even without making them into so-called moral machines.

This chapter has shown here that AMAs are promoted for reasons of inevitability, complexity, establishing public trust, preventing immoral use, because they would be better moral reasoners than us, or because there would be a better understanding of human morality with AMAs. None of these reasons—as they have been articulated in the literature—warrant the development of moral machines, nor will they work in practice. This is so because of: inherent bias to learn how to be ethical, the impossibility or difficulty of understanding the complexity of the robot's decision, how to evaluate or trust the superior ethical reasoning of the robot, and so on.

There are dangers in the language used for these endeavors. One should not refer to moral machines, artificial moral agents, or ethical agents if the goal is really to create safe, reliable machines. Rather, they should be called what they are: safe robots. The best way to avoid this confusion, considering that no critical or unique operational function appears to be gained through

the endowment of ethical reasoning capabilities into robots, is to simply not do it. To that end, the authors suggest an implication for policymakers and academics: place a moratorium on the commercialization of robots claiming to have ethical reasoning skills. This would allow academics to study the issues while at the same time protecting users—the consumer, the indirect user, and society at large—from exposure to this technology, which poses an existential challenge.

In closing, our goal for this article was to pick apart the reasons in favor of moral machines as a way of shifting the burden of proof back to the machine ethicists. It is not up to ethicists anymore to tell you why they think the pursuit of an AMA is flawed; rather, now that it has been shown that the motivations for developing moral machines do not withstand closer inspection, machine ethicists need to provide better reasons. So, to the machine ethicists out there: the ball is in your court.

Chapter 5.

A Misdirected Principle with a Catch

Explicability for AI[49]

5.1 Introduction

It is rare to see large numbers of ethicists, practitioners, journalists, and policy-makers agree on something that should guide the development of technology. Yet, with the principle requiring that artificial intelligence (AI) be explicable, we have precisely that. Microsoft, Google, the World Economic Forum, the draft AI ethics guidelines for the EU commission, etc. all include a principle for AI that falls under the umbrella of 'explicability.' The exact wording varies. Some talk of 'transparency,' others of 'explainability,' and still others of 'understandability.' Finally, Floridi et al.

[49] A version of this was previously published as: Robbins, Scott. 2019. "A Misdirected Principle with a Catch: Explicability for AI." Minds and Machines 29 (4): 495–514. https://doi.org/10.1007/s11023-019-09509-3

call for a principle of 'explicability' for AI, which claims that when systems are powered by AI, humans should be able to obtain "a factual, direct, and clear explanation of the decision-making process" (Floridi et al. 2018).

The intuition that an algorithm should be capable of explaining itself is strong-especially algorithms operating in morally significant contexts. Frank Pasquale's Black Box Society (2015) provides examples of decisions made about us by algorithms for which we are not offered an explanation. It is unfair that we can receive a low credit score, be investigated, be detained, end up on a police watch list, get higher prison sentences, etc. without explanation about the considerations that led to those decisions. If algorithms are used to make decisions in these contexts, there should be explanations about how they arrived at a specific decision.[50] Floridi et al. argue that AI will constrain rather than promote human autonomy unless we have the "knowledge of how AI would act instead of us" (2018, 700).

Getting algorithms to provide us with explanations about how a particular decision was made allows us to keep 'meaningful human control' over the decision. That is, knowing why a particular decision was reached by an algorithm allows us to accept, disregard, challenge, or overrule that decision.[51] 'Meaningful human control' was initially used as a principle for lethal autonomous weapons systems: "humans not computers and their algorithms should ultimately remain in control of, and thus morally responsible for relevant decisions about (lethal) military operations" ("Article 36" 2015, 36).

[50] Robbins and Henschke make the important point that this argument can be turned on its head: "The solution, therefore, is to use such algorithms for specific situations in which it is acceptable to not have an explanation" (Robbins and Henschke 2017).
[51] This is not the only conception of meaningful human control in the literature. More will be said about this in what follows.

'Meaningful human control' is now being used to describe an ideal that all AI should achieve if it is going to operate in morally sensitive contexts (see, e.g., Robbins 2020; Santoni de Sio and van den Hoven 2018). A principle of explicability, then, is a *moral* principle that should help bring us closer to acceptable uses of algorithms. The question then is: does a principle of explicability overcome ethical issues associated with the use of algorithms?

In what follows, I will argue that principles requiring that AI be explicable are misguided. Not only would such a requirement trade off the power of AI in terms of performance, but such a requirement assumes that we have a list of considerations that are acceptable for a given decision. I argue that such a list would preclude the use of machine learning algorithms. Of more philosophical importance is that the property of 'requiring explicability' is incorrectly applied to AI. The real object in need of the property of 'requiring explicability' is the result of the process—not the process itself. This, of course, means that the process itself will need to provide an explanation; however, it only needs to do so if the result requires an explanation. We do not require everyone capable of deciding to be able to explain every decision they make. Rather, we require them to provide explanations when the decisions they have made require explanations. For AI, we should take a similar approach.

Instead of trying to have our cake and eat it too (having powerful AI that can explain its decisions), we should be deciding which decisions require explanations. Knowing that a specific decision requires an explanation (e.g., targeting a person for a drone strike) gives us good reason *not* to use opaque AI (e.g., machine learning) for that decision. Any decision requiring an explanation should not be made by machine learning (ML) algorithms. Automation

is still an option; however, this should be restricted to the old-fashioned kind of automation whereby the considerations are hard-coded into the algorithm. Luckily for the ML community, many decisions benefit society without requiring explanations.

5.2 Calls for a Principle of Explicability for AI

It would be shadowboxing to argue that a principle of explicability for AI is unnecessary if there were no proposals for such a principle. In this section, I highlight some examples of the many calls for such a principle by academics, NGOs, corporations, etc. It should be clear that explicability is considered to be a vital part of achieving the so-called 'ethical,' 'responsible,' 'trustworthy,' etc. AI.

Before highlighting the many examples of calls for a principle of explicability for AI, it is essential to distinguish between the usefulness of explicable AI and a requirement that AI be explicable. I do not argue against the idea that explicable AI could be useful in certain contexts; rather, I will argue against a principle requiring that AI be explicable. For example, if someone were to have an ML algorithm that was highly accurate concerning making predictions about the weather, there may be some desire to have that algorithm explain itself. This desire would not be based on the idea that it is wrong to use the decisions made by the ML without explanation; rather, knowing what considerations were used by the ML for its decision may increase our knowledge about the weather. This example is in contrast with the examples used by those proposing a principle of explicability for AI. ML used for medical diagnosis (de Bruijne 2016; Dhar and Ranganathan 2015; Erickson et al. 2017), judicial sentencing (Berk et al. 2016; Barry-Jester et al. 2015), and predictive policing (Ahmed 2018; Ensign et al. 2017; Joh 2017), and predicting terrorist activities (Mo et al.

2017; Desmarais and Cranmer 2013; Uddin et al. 2020) are just a few of many real-world examples. Using the decisions of ML algorithms in these contexts without explanation is wrong - so the argument goes - unless that ML algorithm is explicable.

One reason that using inexplicable decisions in morally sensitive contexts like the ones listed above is wrong is that we must ensure that the decisions are not based on inappropriate[52] considerations. If a predictive policing algorithm labels people as terrorists and uses their skin color as an important consideration, then we should not be using that algorithm. There could be a case where skin color *is* an empirically sustainable heuristic for determining whether or not someone is a terrorist. For example, in a particular jurisdiction, it may be a matter of fact that all members of terrorist groups are jihadists – and they are overwhelmingly of 'middle eastern appearance.' This could make skin color a legitimate consideration for determining whether or not someone is a terrorist (combined with other considerations of course).[53] If the algorithm is not explicable, then this possibly unethical consideration may be used without our knowledge. The opacity of the algorithm prevents us from knowing whether it is unethically biased.

One of the main reasons that AI, and ML specifically, is the target in calls for a principle of explicability is that these algorithms are opaque. The inputs used for ML algorithms[54] are translated into a machine-readable format (1s and 0s), and then based on the patterns those 1s and 0s, have a path is taken through a series of hidden layers. For simplicity, we can think of these hidden layers as a

[52] Innapropriate captures both considerations that are unethical (e.g. race) and clearly irrelevant (e.g. your astrological sign). Both are innapropriate and could lead to unethical outcomes.
[53] For more detail on the acceptability of profiling see Shauer (2003)
[54] I specifically discuss deep learning algorithms here. Note that other ML algorithms using different methods exist (e.g. evolutionary algorithms).

decision tree for the algorithm. If we were deciding whether or not to target and kill someone, we might ask, "Is this person armed?" and, if yes, "are they pointing their weapon at me?". The algorithm will do something like this with considerations that we cannot understand. The data used to train this algorithm will have given each of the many paths that an input could take a probability corresponding to the resulting classification. Although many researchers are working to make this process explicable, little progress has been made (see, e.g., Gilpin et al. 2018; Kuang 2017; Wachter et al. 2017). Those who have had some success can only give us educated guesses based on many results. In a nutshell, they are using algorithms to analyze the results for patterns that may tell us something about the reasons used by the target ML algorithm.

In short, we do not know the reasons for a specific ML algorithm decision. Combine this fact with using ML algorithms for decisions that have moral significance (i.e., decisions which could result in harms that are rights violations), and we have an ethically problematic situation. An algorithm used, for example, to accept or reject your loan request will significantly affect you. A rejection could cause you and your partner significant distress and change the course of your life. It is precisely this type of situation that motivated the European Union to include in the General Data Protection Regulation (GDPR) what many have interpreted as a 'right to explanation' when fully automated decisions significantly affect someone:

> *the right not to be subject to a decision, which may include a measure, evaluating personal aspects relating to him or her which is based solely on automated processing and which produces legal effects concerning*

him or her or similarly significantly affects him or her[55]

It is intuitive that, when an ML algorithm makes a decision about us that has a morally 'significant' effect, it should be able to 'explain' itself. This intuition has led many to propose that a principle of AI is that it should be explicable. The US Department of Defense believes that explainable AI "will be essential if future warfighters are to understand, appropriately trust, and effectively manage an emerging generation of artificially intelligent machine partners" (Turek n.d.). Below are further calls from academics, non-governmental organizations, and large technology companies for an explicability principle for AI.

Luciano Floridi, for example, outlined a framework for a 'Good AI Society.' In that framework, he and his colleagues explicitly call for AI systems that make 'socially significant decisions' to be explicable:

> *Develop a framework to enhance the explicability of AI systems that make socially significant decisions. Central to this framework is the ability for individuals to obtain a factual, direct, and clear explanation of the decision-making process, especially in the event of unwanted consequences. (Floridi et al., 2018, p. 702)*

NGOs, including the Public Voice (established by the Electronic Privacy Information Center) and the Future of Life Institute, have also called for principles of explainability for AI.[56] The Public Voice, in their list of AI Universal Guidelines, has a right to transparency which states:

[55] GDPR Recital 71. The full text can be found at https://gdpr-info.eu/recitals/no-71/. Some have argued that no such right can be derived (Wachter et al. 2016).

[56] For other examples of principles which could be interpreted as requiring AI to be explainable see UNI Global Union (2018), the Partnership on AI (2019). There are sure to be more.

> *All individuals have the right to know the basis of an AI decision that concerns them. This includes access to the factors, the logic, and techniques that produced the outcome. (The Public Voice 2018)*

And the Future of Life Institute includes two transparency principles in their AI Principles:

> ***Failure Transparency:*** *If an AI system causes harm, it should be possible to ascertain why.*
>
> ***Judicial Transparency:*** *Any involvement by an autonomous system in judicial decision-making should provide a satisfactory explanation auditable by a competent human authority. ("AI Principles" 2017)*[57]

Microsoft's current CEO, Satya Nadella, called for a transparency requirement in an op-ed to the online magazine Slate:

> *A.I. must be transparent: We should be aware of how the technology works and what its rules are. We want not just intelligent machines but intelligible machines. Not artificial intelligence but symbiotic intelligence. The tech will know things about humans, but the humans must know about the machines. People should have an understanding of how the technology sees and analyzes the world. Ethics and design and in hand. (Nadella 2016)*

And Google claims that they will "design AI systems that provide appropriate opportunities for feedback, relevant explanations, and appeal" ("AI at Google: Our Principles" 2018).

Last but not least, James Bridle, in his book *The New Dark Age: Technology and the end of the Future,* calls for a fourth principle of robotics (to add to Asimov's first three): "a robot—or any other intelligent machine—must be able to explain itself to humans" (Bridle 2019).[58]

---

[57] It is unclear why the judicial context gets special attention here. While the judicial context is an especially morally salient one, it is none more so than medical or policing contexts.

[58] It must be noted that Asimov originally had four total laws-meaning that Bridle's would be a fifth, not a fourth. He added a 'zeroeth law'

While it is not possible to claim that this sample of principles, and the many others I did not mention, all amount to the same thing, they do all call for AI to be explicable. To be sure, I do not think it is the intent of the authors of these principles to require *all* AI to be explicable; however, the way that the principles are written, this requirement would either apply to all AI, or it would be unclear when it would have to be applied or not. In some of the examples above, the principles call for transparency of AI. Although transparency and explicability are not synonymous, when transparency is used concerning the transparency of the reasons for the AI-generated decision, this amounts to explicability. Others have called for transparency principles, which are not the same as explicability. Instead, what they mean by transparency is the transparency of the sourcing and usage of training data or transparency of other parts of the development and implementation of AI.[59] One can support this kind of transparency without supporting a principle of explicability. One may, for example, be transparent about the training data used for the algorithm without being able to explain a particular decision made by that algorithm. This kind of transparency would go some way toward ensuring that algorithms will work for a diverse set of people (e.g., ensuring that the training data was not solely made up of the data regarding white males).

The many examples highlighted above are there to make it clear that there are many calls for AI to be explicable. Indeed, not just calls, but demands for a principle that would require AI to be explicable. It is the purpose of this chapter to argue that such a principle is misguided.

---

to precede the others which stated: "a robot may not harm humanity, or, by inaction, allow humanity to come to harm."

[59] See e.g. Whittaker et. Al. (2018)

5.3 The Why, Who, and What of an Explicability Principle for AI

5.3.1 What is the Purpose of Explicability?

Before getting to what explicability is and who it is for, we must understand what the purpose is for a principle of explicability for AI. This will go some way towards understanding what explicability is and who it is for. I argue that a principle of explicability is primarily for the maintaining of meaningful human control over algorithms. The idea is that an explanation of an algorithm's output will allow a human being to have meaningful control over the algorithm–enabling the ascription of moral responsibility to that human being (or set of human beings). With an explanation of the algorithm's decision, human beings can accept, disregard, challenge, or overrule that decision. The Center for a New American Security (CNAS), for example, writes that it is necessary that "human operators are making informed, conscious decisions about the use of weapons" and that "human operators have sufficient information to ensure the lawfulness of the action they are taking…".[60]

There are, however, other features of meaningful human control that would not be captured by explicability. Meaningful human control over autonomous driving systems may not require human beings to have any say over a particular decision because of the psychological limitations of the human driver to gain cognitive awareness in time to act (Heikoop et al. 2019). Santoni de Sio and van den Hoven (2018) argue that meaningful human control occurs when algorithms meet 'track' and 'trace' conditions. We must be able to trace moral responsibility for the outcomes of algorithms back to human beings. The decisions of algorithms must also track

[60] For other documents with similar features for meaningful human control see e.g. United States Department of Defense (2012). For a helpful overview of the common themes involved in discussions about meaningful human control see Ekelhof (2019)

human values. While I use a specific conception of meaningful human control (i.e., giving humans the ability to accept, disregard, challenge, or overrule an AI algorithm's decision), I am not arguing that this conception is the best one. Rather, this is the conception that I argue is implicit when one requires that AI be explicable.

We must keep in mind that an explicability principle for AI is ethical. The starting point for these lists is that there are ethical problems associated with algorithms. If the design and development of algorithms follow a particular set of principles, then it is believed that the resulting practice using the algorithm will be 'good,' 'trustworthy,' or 'responsible.'[61] So, a principle of explicability is an attempt to overcome some ethical issues unique to algorithms.

Ethical value is to be contrasted to the epistemic value explicable AI might provide. Explicable AI may be extremely valuable to researchers and others who would be able to use explanations to understand their domain better. Garry Kasparov, for example, may find an explanation of a particular chess move made by an algorithm beneficial for his ability to play chess.[62] A doctor may find an explanation useful to understand better how to diagnosis a particular disease. This epistemic value of explicability for AI is not under dispute. In these cases, we are not harmed by the opacity of the algorithm's decision-making process. A principle of explicability, in contrast, is ethical in that it is about

[61] Using the terms 'good', 'trustworthy', or 'responsible' in relation to AI can confuse people into believing that those adjectives refer to the algorithm itself – yet no algorithm can be 'responsible' in a moral sense. What is really referred to here is the practice of using AI – not the algorithm itself.

[62] A good, recent, example of this is the growing discussion about Move 37 by AlphaGo during its game with Lee Sedol (Metz 2016).

preventing harm (broadly construed) that could occur due to the opacity of the algorithm.

What is the ethical issue that is giving rise to this principle? One candidate is the issue of understanding what went wrong if something harmful happens as a consequence of the algorithm. For example, if an autonomous weapon used against terrorists struck a wedding party and killed 100 civilians, then it would be helpful to have an explanation of what caused this to happen to prevent it from happening in the future. While a principle of explainability would help with this, it does not capture the full range of ethical issues that explicability aims to overcome. For example, if someone is incorrectly placed on a terrorism watch list by an algorithm, how will we know that something harmful has happened so that we can demand an explanation?

This points to the ethical issue of ensuring that the outputs of algorithms are not made based upon ethically problematic or irrelevant considerations. We expect, for example, placement on a terrorism watch list not to be based on the color of the applicant's skin (or a proxy thereof). An explanation of the algorithm's decision can allow for someone to accept, disregard, challenge, or overrule the rejection. This gives meaningful control of the decision to human beings. An explanation of the algorithm's output goes above and beyond the stipulation that some particular human is responsible for the algorithm's decisions. It provides a human with the information they need to exercise that control.

Explicability, therefore, is an attempt to maintain meaningful human control over algorithms. Only human beings can be held morally accountable, so it should be human beings that are in control over these decisions (see, e.g., Johnson 2006b). If a human being has an explanation of the algorithm's decision, then it is

possible for that human being to accept, disregard, challenge, or overrule that decision.

5.3.2 Who is Explicability for?

How the requirement that AI be explicable is understood depends upon who will receive the explanation. A medical diagnosis algorithm that classifies someone as having a brain tumor might, for example, provide a heat map of which parts of the brain scan most contributed to the diagnosis. This 'explanation' would probably be useless to a patient–or anyone else without very specific medical training. However, if the goal is that the algorithm is under 'meaningful human control,' then we are not concerned with the patient's understanding of the explanation.

Just as with any diagnosis, we trust that our physician is making a justified decision in line with current medical practice. The physician should be ultimately responsible for the brain tumor diagnosis, and therefore it is the physician who should be able to evaluate the explanation. In the case of an autonomous weapon targeting someone for killing, it would be useless to provide the person to be killed with an explanation as to why they were targeted. Someone has to be in an epistemic position to be able to reject or endorse what that autonomous weapon is doing. Does this explanation meet the rules and laws of targeted killing? Remember that the purpose of the explanation is to overcome an ethical problem, namely, to establish meaningful human control over that decision by allowing one to confirm that the reasons for a decision are in line with domain-specific norms and best practices.

To illustrate, let us say that an algorithm rejects a loan application. This algorithm can provide an explanation in the form of considerations that played a factor in its rejection. One of those considerations was the fact that the application included a high debt-to-income ratio. To the applicant, this is interesting to know. Still, it

would be quite unclear whether their debt-to-income ratio was at a level that justified its factoring in on a decision to reject their loan application. Only those with relevant domain-specific knowledge would be able to evaluate whether this particular debt-to-income ratio should factor into a decision to reject the loan. This only gets more complicated as more considerations factor into algorithmic decisions.

To achieve the ethical goal of a principle of explicability, the explanation provided by an algorithm should enable a human being to have meaningful control over the decisions the algorithm makes. This means that the person using the algorithm is the person that the explanation should be directed towards–not the person subject to the decision of the algorithm (although those two roles may be filled by the same person). While the person subject to the algorithm's outputs may be interested to know the explanation (and in some cases should be provided with it to achieve other ethical goals[63]), this does not establish meaningful human control over the algorithm's output.

5.3.3 Artificial Intelligence

'Artificial Intelligence' is an overused phrase that signifies many things. Explanation also has many uses depending on the context. We have had artificially intelligent systems for decades that did not result in any calls for explanation. This is mainly because what is known as good old-fashioned AI (GOFAI) is simply a set of explicitly coded rules in the form of a decision tree that allows for the automation of processes. For example, if you wanted to automate the decision on which move to make in chess, it may look like this:

[63] Most notably the goal of *actionable recourse*: the ability to contest incorrect decisions or to understand what could be changed in order for the data subject to achieve a more desirable result (Wachter et al. 2017; Ustun et al. 2019).

```
If (first move of game) then move random pawn 2
      spaces forward

Else if (king is in check) then (move king to non-
      checked space)

Else if (possible to achieve checkmate) then
      (achieve checkmate)

Else if (possible to put king in check) then (put
      king in check)

Else if (possible to take an opposing players piece)
      then (take piece)

Else (move piece at random)
```

This is a terrible algorithm for deciding your next chess move. A much more sophisticated algorithm designed using GOFAI could be achieved. However, this kind of automation is inherently explicable because the code makes the reasons for a resulting decision explicit. Opacity concerning this type of automation would only occur if the institutions doing the automating did not want people to know how the decisions are being made (see, e.g., Pasquale 2015).

GOFAI is in contrast to AI that falls under the umbrella of machine learning (ML). The GOFAI approach is limited by what considerations the designers of the algorithm could think of to incorporate into the decision tree. Novel situations may result in terrible decisions by the AI. ML is one approach to overcome such limitations. In a nutshell, ML attempts to use statistical methods to allow an algorithm to 'learn' every time it 'tries' to achieve its specified goal. Each attempt, whether it fails or succeed, will result in the algorithm updating its statistical probabilities that correlate to features of the input.[64]

[64] For a nice overview of machine learning methods and trends see Jordan and Mitchell (2015).

An ML algorithm could be trained to play chess by playing many times without explicit rules given by humans. The ML algorithm may play at random the first time–losing very easily. At the end of the game, we would tell the AI that it lost. In the next game, the AI would play slightly differently. Over hundreds, thousands, or even millions of games, the AI would be very well trained to play the game of chess. The resulting trained ML algorithm would be opaque concerning its reasoning for any given move.

Is it acceptable that the algorithm makes decisions that are not explicable? If you share my intuition that there is no problem here, it may stem from the fact that the outcomes of these 'chess move' decisions cannot result in harm. A terrible chess move may result in the loss of the chess game, but life, limb, reputation, and property are not at stake. An AI making decisions in other contexts, such as medical diagnosis and judicial sentencing, could cause real harm.

The point here is to show that the principle of explicability is vital due to the rise of algorithms using ML or other methods that are opaque with regard to how the algorithm reaches a particular decision. If we are simply using automated processes (e.g., GOFAI), then explicability is only a problem if the developer intentionally obfuscates the explanation. In these cases, an explanation is readily available to developers and companies; however, they do not see it in their interest to reveal that explanation to the public. While not addressed here, this problem is critical (see Pasquale 2015).

5.3.4 Explicability

So if one is using an ML algorithm for decisions that could result in harm and responsibly wants to adhere to a set of principles that includes a principle of explicability, what is one to do? First, one would need to know what is being demanded by a principle of

explicability. That is, what is an explanation that would satisfy the principle?

First, we could be demanding a causal explanation for a particular outcome/action/decision. For example, when Google's image classification algorithm classified two young black people as gorillas, there was an outcry and much embarrassment for Google (Kasperkevic 2015). Suppose Google was to explain the algorithm's classification by saying that "features of the image input correlated highly with training images classified as gorillas," I doubt that anyone would be satisfied. We are not concerned with how the algorithm classifies images in general. Rather, we want to know why the label 'gorillas' was applied to a specific image by the algorithm. In other words, we demand to know the specific features of the image that contributed to the labeling.

Scientific explanations also give us answers to *how* things happened. However, we do not want to know the *how*; rather, we want to know the *why*. I do not want to know how my daughter hit her brother: "I raised my right arm and moved it forward at high velocity," but the why: "he took my favorite stuffed animal from me." The latter *why* explanation is an explanation that provides the reason(s) that a particular action was taken. This reason or reasons may or may not morally justify the action. These reasons are precisely what we want to evaluate. Some reasons will be good reasons – they justify the output. Some reasons will be bad reasons in that they fail to justify the output. This could be due to the fact that the reason used is unethical (e.g., labeling someone a criminal due to their race) or because the reason is irrelevant (e.g., labeling someone a criminal because they have three vowels in their name). In the case of ML, we could get an explanation like the following excerpt used to describe how DeepMind's AlphaGo chooses its next move:

> *At the end of the simulation, the action values and visit counts of all traversed edges are updated. Each edge accumulates the visit count and mean evaluation of all simulations passing through that edge is the leaf node from the ith simulation, and 1(s, a, i) indicates whether an edge (s, a) was traversed during the ith simulation. Once the search is complete, the algorithm chooses the most visited move from the root position (Silver et al. 2016)*

This, if you are a person with the requisite knowledge to understand it, is an explanation of the *how* for a particular move in the game of Go made by the algorithm-driven process. It says nothing about the particular features of that move, which contributed to the decision to make the move. That is, nothing in that explanation gives us reasons which justify that particular move. One could attempt to justify a particular move made by the algorithm by referencing the effectiveness of the algorithm itself: "the move chosen by the algorithm is a good move because the algorithm has proven to be very good at the game of Go." We can see that this is an unsatisfying explanation when we apply it to a different context. If a great sniper (who has yet to kill a civilian) were to blow the head off of a young child and her superior were to ask: "why did she kill that young child?" and someone were to respond with "it was good to kill the child because the sniper has never killed a civilian before" we would not, and should not, be satisfied. What we want with an explanation are all (and only) the considerations necessary for their contribution to a particular decision — considerations that a human could use to determine whether a particular algorithmic decision was justified.

We could give a general explanation of sorts for opaque algorithms in any context. Why did the ML algorithm decide to label a convicted terrorist as high-risk, i.e., as likely to re-offend? Because data used as an input to the algorithm correlated with features of data used to train the algorithm that was tagged as having a high risk. While

this is an explanation, it falls short of what is desired by the principles highlighted above. What is desired is an explanation that would provide a human with information that could be used to determine whether the result of the algorithm was *justified*.

An explanation may justify a particular decision, or it may not, and a decision may be justified by reasons that do not feature in an explanation of that decision (see, e.g., Dancy 2004, chap. 5; Darwall 2003). If, for example, I were to make a move in chess because I thought that it would make the board more balanced (in terms of aesthetics), we would have an explanation for the move that I made that failed to justify the move. However, that move may also have been the best move I could have made – making the move justified. While it was a great chess move, I doubt anyone would take my advice on a future move - nor should we if we knew that an algorithm was using board balance as a consideration in favor of a particular move. This shows that we cannot merely look to the decision itself and ask whether that decision was justified or not. An algorithm may flag someone as a dangerous criminal who happens to be a dangerous criminal – justifying the algorithm's classification. However, if the consideration leading to that classification was the person's race, then we have an explanation that fails to justify the decision whether the decision was correct.

In short, what is desired is an explanation providing the considerations that contributed to the result in question. This gives a human being the information needed to accept, disregard, challenge, or overrule the decision. In the same way that a police officer might claim in court that a particular person who committed a terrorist is at high-risk of reoffending, and the judge asks for the considerations used to justify such a label, we want the algorithm to justify itself by telling us what considerations were used.

A justification for this 'high risk' label given by the police officer might be that while in custody, the terrorist reaffirmed her commitment to the terrorist group and its strategy of killing innocent persons; indeed, she threatened to conduct terrorist attacks if she was free. The judge may accept this as a good justification and sentence the criminal to the maximum allowable prison sentence. If, on the other hand, the police officer justified this label by saying that the terrorist was dark-skinned and menacing looking, then the judge (hopefully) would reject the police officer's label of 'high-risk.' If an algorithm were delegated the task of labeling criminals as 'high-risk' and did so as a result of race, then we would want the judge to know that so that she could reject the algorithm's decision. A technical, causal, or scientific explanation does not allow the judge to have meaningful human control over the algorithm.

5.4 Current Approaches to Explicable AI

Having a principle requiring that AI be able to explicable means that there must be methods for which an algorithm can give an explanation for its decision. Here I do not focus on intrinsically explainable algorithms (like the GOFAI approach above). Instead, I focus on the ML algorithms that are the reason for introducing a principle of explicability for AI in the first place.

There has been much work in achieving explainable AI. This work can be classified into two broad approaches. The first is offers 'model-centric explanations', and the second offers 'subject-centric explanations' (Edwards and Veale 2017, 22). Model-centric explanations aim to provide the information that is known about the algorithm to understand the algorithm better–enabling users to understand better how to use the algorithm. The information that provided relates to the data the algorithm was trained on, how the algorithm was tested for bias, the intentions of the designers, performance

metrics, etc. The idea is that knowing all of this other information about the algorithm may allow society to "make informed choices regarding usage, implementation, and regulation of these machines" (Robbins 2020).

While this approach to explainable AI is interesting, it does not capture what is meant by 'explicability.' We do not have the considerations that played a factor in the resulting decision. At best, we have guestimates, or maybe a justified belief that the algorithm will work in the given context because it has performed well in similar contexts, and the input is relevantly similar to data used during the training phase of the algorithm. This does not overcome the ethical problem resulting in important decisions made by algorithms. However, it may significantly help society decide on the acceptability of using a specific algorithm for a specific purpose.

The 'subject-centric' explanations are an attempt to zoom in on the input (the subject) and understand what it is about it that caused the specified decision. For example, an explanation of a loan rejection may be that the person who requested the loan has a debt-to-income ratio that is always classified as a rejection by the algorithm. While there may be other considerations that would also contribute to a rejection, the debt-to-income ratio could be seen as a sufficient condition for rejection. In this clear cut case, the explanation would help humans decide whether the decision was justified–and therefore satisfy the type of explanation discussed in the previous section. Unfortunately, ML decisions are rarely going to be this simple. The more data fed into the algorithm as input makes the output that much harder to explain. Many variables may need to be modified to change the resulting classification–making it increasingly unlikely that a satisfactory explanation is provided.

5.5 Three Misgivings about Explicable AI

There are three major misgivings I have regarding the principle of explicability for AI. The first is with regard to where the property of 'requiring explicability' is placed. I argue that we do not normally place such a property on the process, which results in a decision; rather, we place that property onto the decision itself. Second, there seem to be many implementations of AI in situations of low to no risk (in terms of harm). It is unreasonable that the decisions resulting from AI in these situations should be required to provide explanations. Finally, in situations of high risk, there is a catch-22 for those who wish to use ML: If ML is being used for a decision requiring an explanation, then it must be explicable AI, and a human must be able to check that the considerations used are acceptable, but if we already know which considerations should be used for a decision, then we don't need ML.

5.5.1 Explicability of the Decision vs. Decision-Maker

The mistake with requiring that AI be explicable is that it places the requirement of explicability onto the decision-maker rather than the decision itself. Some calls for a principle of explicability allude to this when they add the qualifier resembling 'when the decision made by the AI significantly affects a person.' This is an acknowledgment that the property of 'requiring an explanation' really applies to the decision itself—not the entity making that decision.

When my daughter hits her brother, I would reasonably demand that she explain her decision to act in that way. She has significantly impacted her brother because she has directly caused him pain. In contrast, when my daughter suddenly starts to dance, and I ask her why, she would (and has done many times) shrug her shoulders and say, "I don't know." I, of course, am not mad at her for her lack of explanation. The reason is that one action requires an

explanation, and the other does not. The first action resulted in harm, thereby 'significantly affecting' a person. The second action is benign. No one is harmed by my daughter's spontaneous dancing. It would, therefore, be unreasonable if I were to tell my daughter that everything she did requires a morally justifying explanation[65] or that all children should be 'explicable.'

In short, adding the property 'requiring explicability' to children would be a mistake. It is the action or decision which can/should have the property of requiring explicability. Decisions capable of causing harm (broadly construed) are decisions that require this property. Anyone unable to give an explanation for such a decision is doing wrong.

When discussions about AI and explanation come up, there are some common examples given. Algorithms making decisions about loan applications, criminal sentencing, policing, medical diagnoses, weapons targeting, etc. all get mentioned when discussing the need for algorithms to be able to explain themselves. However, the common element in all of these contexts is that the decisions made in these contexts require explanations that justify those decisions. Whatever the process used to make these decisions, there must be a justifying explanation for any given decision.

This is important because using explicability as a principle for AI could force those designing algorithms for decisions or roles that do not require explanation to use less powerful AI like GOFAI. This would significantly constrain many of the great uses of ML algorithms that are not able to explain themselves. For example, ML is often used for credit card fraud detection (Morrell 2018). When

[65] This does not preclude my interest in an explanation in terms of, for example, her desires and preferences. If she simply told me that she "loved dancing" when asked "why" then this may provide me with a reason for entering her into dance class.

the algorithm classifies a transaction as fraudulent, this causes the bank to lock the credit card until the customer can confirm that they indeed made the transaction. False positives can, to be sure, be annoying; however, the only thing we care about is whether the algorithm performs well compared to other methods. Because the role of the algorithm is simply to flag a transaction as fraudulent, the ultimate decision-maker will be the customer herself. I can see no good reason why the ML algorithm should be forced to provide an explanation here.

This is why many of the principles highlighted above include a qualification, namely, that AI must be explicable if the decision will significantly affect someone. This, of course, needs to be specified very clearly to separate the decisions that will trigger the principle and those that do not. Of course, once we do this with any level of specification, we are simply deciding what roles, tasks, and decisions require explanations and which ones do not. The principle will no longer have anything to do with artificial intelligence.

5.5.2 Inexplicable AI for Low-Risk Purposes

In May 2015, Google's AlphaGo algorithm defeated the world champion Go player Ke Jie (France-Presse 2017). The AlphaGo algorithm provided Ke Jie no explanation for any of the moves it made. However, an algorithm deciding which moves to make in the game of Go does not seem problematic because the possible consequences stemming from these decisions are at no risk of causing harm. Many AI and ML applications fall into this category. This is more often than not a result of the algorithm's implementation within a larger process. For example, Cortis is an algorithm that detects voice patterns associated with cardiac arrest (Vincent 2018). The algorithm exists explicitly to aid emergency call operators. The algorithm takes as its input live sound from the calling line. Its output is true if the voice pattern is associated with cardiac arrest and

false if it is not. The context of emergency calls is high risk. The operator has legal, as well as moral, responsibility, and can make decisions that will save (or end) lives. The addition of the algorithm in this context aids the operator with one specific problem: someone on the other end of the line may be having a heart attack.

This algorithm cannot, however, cause harm. The worst-case scenario is that the algorithm does not identify someone as having a cardiac arrest who is indeed experiencing cardiac arrest. This is regrettable; however, the algorithm not being there would not have changed this outcome.[66] It is an example of an algorithm that should be judged on its accuracy–not its reasons. A principle of explicability would mean this algorithm would not be allowed to operate. This would be unfortunate as it has been shown to detect heart attacks on average 30 seconds faster than human operators with an accuracy of 93% (human operators have a 73% accuracy rate).

It should be noted that establishing that a particular algorithm has no risk of causing harm would be incredibly difficult to establish. It will often be the case that it is unknown what the possible consequences of algorithmic decisions will be. There would have to be some standard of risk for automated decisions before we allow anyone to claim that their algorithm's decisions cannot cause harm. The point here is to show that there are cases where algorithm's decisions have a low risk of causing harm, and a lack of explanation should not preclude its use.[67]

5.5.3 Catch 22 of Requiring Explicability for AI

In Joseph Heller's Catch 22, Doc Daneeka explains to Yossarian the catch regarding the policy allowing insane

[66] There is a concern that operators may come to think that there is no heart attack unless the algorithm identifies one – resulting in situations where were there not an algorithm they would have identified a heart attack on their own (thanks to Prof. Seumas Miller for this concern).

[67] Thanks to an anonymous reviewer for making this point

people to cease flying bombing missions: "Catch 22. Anyone who wants to get out of combat duty isn't really crazy" (Heller 2011, 52). So to get out of combat duty, one would have to be insane and to tell their superior that they wished to cease combat duty. Unfortunately, only a sane person would make such a request. There is a similar catch to explainable AI. If ML is being used for a decision requiring an explanation, then it must be explicable AI, and a human must be able to check that the considerations used are acceptable. But if we already know which considerations should be used for a decision, then we do not need ML.

An example may help to illustrate: say there is an ML algorithm that is developed to decide whether someone should be placed on a terrorism watch list. This algorithm is opaque, and there are justifiably calls for explainable AI in this context. So we pour millions in funding to come up with explainable AI that somehow is just as powerful as the original algorithm.[68] Now when someone is placed on the terrorism watch list, there is an explanation spit out by the algorithm. A human analyst can check this explanation to ensure that it is an 'acceptable' explanation – i.e., that it does not include irrelevant or unethical considerations. We can imagine an explanation for this being "the person is Muslim, resides in a poor neighborhood, and likes the movie the Godfather." This is a terrible explanation that does not justify the placement on a terrorism watch list. The inclusion of the factors 'Muslim' and 'poor neighborhood' could, if combined with many other factors, justify the decision. However, on their own, they simply single out poor Muslims, which does not justify their reduced autonomy. Furthermore, the explanation includes the completely irrelevant factor

[68] This is unlikely as there is widespread acknowledgement that explainability and power conflict and must be traded off in the context of AI.

about liking the movie the Godfather. This decision by the algorithm should be rejected.

On the other hand, we can imagine a decision and explanation by the ML for placing someone on a terrorism watch list that is 'acceptable.' The explanation might be something like "the person frequently uses the term jihad on online discussion forums, has downloaded issues of *Inspire* magazine, and frequently communicates over the encrypted Telegram platform." These considerations, taken together, may justify this particular person's inclusion on a terrorism watch list. The problem with all of this is that for the explanation given by explainable AI to be useful, we must have a human capable of knowing which considerations are acceptable and which are not. If we already know which considerations are acceptable, then there is no reason to use ML in the first place. We could simply hard-code the considerations into an algorithm–giving us an automated decision using pre-approved, transparent reasoning.

For an explanation of a decision made by an ML algorithm to be useful, we already need to know what counts as an acceptable consideration for that decision. For example, we can imagine an ML algorithm that could make a modern painting and could give us an explanation for each brushstroke. Since there is no agreed-upon list of considerations that 'justify' a brush stroke in the context of modern painting, it would be a useless explanation. We could do nothing with that explanation concerning the decision it made (e.g., reject its decision). Here, the reader may think that the explanation would still be useful. We may just be curious to know why the algorithm did what it did. Furthermore, if one was a modernist painter, then this information could be used to help them become a better painter. And in the terrorism watch list example, the explanation could point to previously not thought of considerations.

There is no doubt some truth to this. Explainable AI could be used to find correlations that should serve as considerations regarding the class of decisions at hand. However, explicability in these scenarios is very different. Now the explanations proffered by explainable AI are not justifying explanations–they cannot be used to justify a specific decision. For example, if the terrorism watch list algorithm used a consideration that the person looked up a specific material that was to date not on a list of materials that were of concern by the intelligence community, then the proper thing to do would be to find out if such a material was indeed cause for concern. It may be found out that this material has the potential to be used for bomb-making. The explanation, then, may help find considerations that are relevant but not known before. This consideration can now be used by GOFAI or counter-terrorism analysts to place people on watch lists. In cases like these, we would no longer be checking an algorithm's explanation to ensure that it conforms to our view of what's acceptable; rather, the explanation would hopefully point us towards acceptable considerations we hadn't thought of before. Once we have these new considerations, then we could just hard code them into traditional automation algorithms (e.g., GOFAI) rather than let the ML algorithm take the role of decision-maker.

5.6 Conclusion

If my arguments in this article are on the right track, then we will find the solution for the opacity of ML by using ML for roles, decisions, or actions which do not have the property of 'requiring an explanation.' This solution may seem, at first glance, to restrict ML to playing games. If games are the only things without the property of 'requiring an explanation' that ML can do well, then this would be true. However, ML has had much success to date in contexts–like healthcare–that have ethical and societal import. Much of this success has been

making decisions that do not require explanations. Detecting cancerous moles is one such example. An algorithm can take a picture of a mole and classify it as malignant or not. The consequences of this decision are simply a biopsy if the mole is labeled as malignant. That is, there is an independent way of verifying whether or not the algorithm was correct – thereby precluding the need for an explanation. This algorithm also outperforms dermatologists at such classification (Esteva et al. 2017; Presse 2018). The initial classification by a doctor is done by merely looking at the mole–and although there are certain 'rules of thumb' regarding size, color, and shape, it is difficult to articulate what malignant moles look like. A doctor is not required to explain their decision. An algorithm should not be required to either–especially when it outperforms human beings at the task.

One difficulty that arises with algorithms that perform tasks like the one above is that they may still be biased and indirectly harm a group of people. Although it seems that the algorithm has a net benefit to society in that it outperforms doctors at labeling moles malignant–this benefit may not be the same for all groups of people. In this case, the algorithm performs poorly on those with a dark complexion (Lashbrook 2018). Note that this does not have anything to do with explicability as used in principles for AI. The algorithm is not using skin color as a consideration for determining whether a mole is malignant; instead, the algorithm is not very good at labeling moles on patients with dark complexions. To take a more straightforward example, when an individual practices a presentation before a conference, they may be able to pace the presentation well, speak clearly, and not lose their place. When it comes to the actual presentation in front of a group of people, they could still perform much worse. They speak too fast - causing them to end too early–and lose their place, which causes them to skip over slides because they cannot remember what they were

supposed to say for them. They did not decide to perform poorly because they were in front of a group of people. Quite the contrary–they made a conscious effort to perform their best. They are just not very good at presenting in front of people. The source of their problem–and the problem with many ML algorithms–is not in the explanation of the decision but in the efficacy of its decisions/actions given different contexts and inputs.

In this article, I have argued that the property of 'requiring an explanation' belongs to the decisions and actions themselves–not the entity performing the action or decision. When we direct our attention to those decisions and actions, we can better decide in which contexts and roles we should be using ML algorithms. Furthermore, in showing that there is a catch 22 for explicable ML algorithms, it is argued that the reason for making explicable AI is an epistemic one–not a moral obligation. The only way to use explicable ML to solve the moral issue of algorithmic opacity is if we have already figured out the acceptable considerations for making the decision or performing the action at hand. If we already have those acceptable considerations, there is no need to use ML in the first place.

Chapter 6.

AI & the Path to Envelopment

Knowledge as a first step towards the responsible regulation and use of AI-powered machines[69]

6.1 Introduction

Artificial intelligence (AI) and robotics are increasingly entering our lives - from smart assistants in the home, social robots in the hospital, to algorithms delivering our news. There is no shortage of proposals for algorithms and robots in the future to take on novel roles – from AI-powered sex robots (Sharkey et al. 2017) to AI therapy bots (Gaggioli 2017). If implemented responsibly, these algorithms will no doubt positively contribute to society. However, each of these applications brings with it the possibility of new ethical issues or to exacerbate existing ones. Autonomous weapons systems, for example,

[69] A version of this was previously published as: Robbins, Scott. 2020. "AI and the Path to Envelopment: Knowledge as a First Step towards the Responsible Regulation and Use of AI-Powered Machines." AI & SOCIETY 35 (2): 391–400. https://doi.org/10.1007/s00146-019-00891-1

are given morally significant roles. Some have argued that machines like these require moral reasoning capabilities to navigate the ethical dilemmas they are sure to face (Wallach and Allen 2010; Scheutz 2016; Arkin 2008; Arkin et al. 2012). This raises issues regarding the moral status of the machine and of assigning moral responsibility when bad outcomes occur (Johnson 2006b; Bryson 2010a; van Wynsberghe and Robbins 2019). A problem society currently faces is one in which we do not have ethical norms, regulations, or policy guidelines to assist developers in getting the right balance between harnessing the power of AI while at the same time avoiding negative ethical and societal impacts. The first step to solving this problem, however, requires closing an epistemic gap, i.e., society does not know for sure what these algorithms do nor how they were created. Before we can create sound regulation and policy to guide AI development, there must be made available specific knowledge of the products and services powered by AI algorithms. This article aims to start us down a path that will lead us out of the epistemic darkness concerning AI-powered machines.

Much of the focus in AI ethics has been on the opacity of AI algorithms in their 'decision'-making. It is not currently possible to know the reasons for a particular 'decision' or output reached by an AI algorithm.[70] In some cases (e.g., playing chess), this may be a perfectly acceptable situation; however, if the algorithm produces an output regarding whether someone will be targeted for a drone strike, then it has been argued that the situation is unacceptable (see, e.g., Sharkey 2011). The likely lethal outcome resulting from the algorithm warrants an explanation regarding how the algorithm reached the 'decision' or output that a person should be killed. There

[70] Here I am discussing those AI algorithms falling under the umbrella of 'machine learning'. There is work to try and overcome this opacity (see e.g. Wachter et al. 2017; Gilpin et al. 2018); however, nothing so far can give us the specific reasons used to make a particular decision.

are many cases, though, when it would be counterproductive to require such an explanation - we use AI in some cases *because* it can produce outputs based on reasons that are not explainable in human language. This access to a broader range of considerations than just the ones that humans can understand gives AI its power. Given this, requiring it to be explicable in human reasoning terms for its decision may undermine its effectiveness. For example, while an algorithm to detect weapons in the bags of terrorists during airport screening is operating in a morally significant context, the algorithm works well precisely because it doesn't use human articulable reasons for its classification.[71] That is, it is unknown how the algorithm comes to classify a particular bag as having a weapon. However, in this case, it seems acceptable for an algorithm to aid security professionals in finding weapons (assuming it generally gets the correct answer and, thereby, prevents a terrorist attack). Why is it that, in some cases, algorithms with opaque reasoning are acceptable, while in others not?

This tension surrounding algorithmic opacity described above is the inspiration for this chapter. I argue that opaque algorithms are acceptable when they are *enveloped*.[72] The central idea of *envelopment* is that machines are successful when they are inside an 'envelope.' This envelope constrains the system in a manner of speaking, allowing it to achieve the desired output given limited capacities. However, to create an envelope for any given AI-powered machine, we must have some basic knowledge of that machine – knowledge that we often lack.

[71] Commercial products providing this very service can be readily found. See https://www.smithsdetection.com/products/icmore/ for an example.

[72] The term 'envelopement' comes from the robot ethics literature. See e.g. Luciano Floridi (2011a) for a discussion of envelopment in which it is argued that envelopment describes the conditions under which robots would be successful.

The knowledge that we need to create such envelopes is knowledge of the: inputs, outputs, function, boundaries, and training data of the AI. In the case of the AI-powered weapon detection at airports, we know about the: inputs (x-ray scans of bags), the training data (lots of pictures of bags with weapons), the function (to detect weapons), and the outputs (weapon or no weapon). We do not know how it decides to classify the baggage scans, but with all of this other knowledge, an explanation is not needed. Even when we know very little about many of these aspects, it can be acceptable to use given that we constrain the AI appropriately. For example, if an AI-powered machine has the function of exploding improvised explosive devices (IEDs) used by terrorists and we are ignorant about what the possible inputs could be, how it comes to a particular output, and what training data was used, it would be simple to decide that we should use this machine only when there aren't people around who could be harmed by exploding devices. By lowering the possibility of harm by limiting the operating environment of the machine, the possibility of realizing the machine's benefits (locating and detonating IEDs) can be still be realized.

The importance of this knowledge becomes especially salient when the outputs of an AI-powered machine have the potential to be harmful. Here harm is to be understood not only as physical harm but also harms like invasions of privacy, financial harms, and restrictions on autonomy. It is also important to note here that these harms are understood to result from the AI – not the companies irresponsibly collecting data on users of their products. While companies collecting data on their customers may involve several harms – the focus of this chapter is on harms that are the result of artificial intelligence. In theory, an AI digital assistant like the Amazon Echo could operate without Amazon violating users' privacy (Amazon might violate users' privacy if Amazon gave police departments access to the recorded audio). That is, the

device could function properly without Amazon collecting, storing, repurposing, or selling users data. The device could simply process the command and delete the audio. The focus here is restricted to those harms that are possible due to the functioning of the AI (both intended and unintended). When harms like this are present, we must know as much as we can about the properties highlighted above to make informed choices regarding usage, implementation, and regulation of these machines.

This chapter begins by going into more detail on the subject of opacity as it relates to applications of AI. Following this is a discussion of the concept of envelopment as it offers what I argue to be a better solution to AI's opacity problem. This is because many features outside of the inner workings of the algorithm remain opaque to us as well. I argue that enveloped AI will help us regulate, use, and be bystanders to AI-powered machines without the need for so-called 'explainable' AI. Bystanders to AI-powered machines are those people who are forced to engage with them in some way. For example, people biking to work may come across an autonomous car and not know how to act around it. I include users and bystanders because regulation is one part of an overall picture that will guide the responsible introduction of AI-powered machines into society. The people implementing and using these machines must do so responsibly, and the people who are being processed by or are bystanders to these machines must be able to navigate this AI-augmented world ethically. Section 4 delves into the properties that I argue are needed to envelop any given machine properly. Before concluding, I briefly respond to some possible objections and limitations of the proper envelopment of AI-powered machines.

6.2 Opacity and Algorithms

There is much discussion about a lack of transparency when it comes to algorithms. Frank Pasquale argues that we live

in a 'black box' society (Citron and Pasquale 2014; Pasquale 2015). Decisions are made by algorithms that affect many facets of our lives. Many of the stories in the media regarding contemporary AI are about algorithms that fall under the umbrella of machine learning. Machine learning algorithms use statistics and probability to 'learn' from large datasets. The complexity of the statistics involved and what those statistics refer to to 'learn' has led to a situation in which we do not know how these algorithms generate their outputs.

This can be quite disconcerting – and probably unethical in many circumstances. A decision about who gets a loan or not or what length of sentence is given to convicted criminals seems to require reasons. The same can be said about decisions regarding who is placed on the terrorist No-Fly list (Robbins and Henschke 2017). Finding out you are on the No-Fly list or were denied a loan without explanation is arbitrary and unacceptable. The European Union's recent General Data Protection Regulation (GDPR) legislation has been interpreted to include a "right to explanation":

> *the right not to be subject to a decision, which may include a measure, evaluating personal aspects relating to him or her which is based solely on automated processing and which produces legal effects concerning him or her or similarly significantly affects him or her*[73]

Although some disagree that such a right can be derived from GDPR (Wachter et al. 2016), the debate is illustrative of the desire for such a right. While I am not opposed to such a right, I find that it has focused the discussion on how to open up the black box of machine learning algorithms rather than simply bar such algorithms from making such decisions. The question is, "how can these algorithms explain themselves" rather than "what

[73] GDPR Recital 71. The full text can be found at https://gdpr-info.eu/recitals/no-71/

decisions are acceptable to delegate to an opaque machine and its outputs?". The difficulty in answering this second question can be found in our ignorance concerning the basics of many AI-powered machines.

This is because we are in the dark regarding AI. By 'we' I mean consumers, policymakers, lawyers, and academics. By 'in the dark' I mean that we have a general lack of knowledge and understanding about the technology. Take the recent example of the Amazon Echo. In May 2018, a woman reported that an Amazon Echo recorded a private conversation between her and her husband and sent it to one of her husband's employees – all without their knowledge (Chokshi 2018). While it is still unclear exactly how this occurred, Amazon's explanation is disconcerting:

> *As the woman, identified only as Danielle, chatted away with her husband, the device's virtual assistant, Alexa, mistakenly heard a series of requests and commands to send the recording as a voice message to one of the husband's employees. (Chokshi, 2018)*

This explanation of the event offers other consumers who have purchased Amazon's Echo devices with little information regarding how to prevent this from happening to them as well. Consumers (and Amazon) don't understand the combination of sounds that served as inputs into this AI-powered device. Consumers don't know what its boundaries are concerning what it can do. Being an internet-connected device with access to your files, contacts, emails, documents, etc., it seems that there are no virtual boundaries for this device. Consumers also do not know what the functions of this device are. It is presented as an assistant with unlimited capabilities, and its slogan is "just ask." Its outputs include: turning on lights, providing information, reading bedtime stories, ordering products, sending emails, chatting, ordering an Uber, etc. There are lists online detailing what possible

outputs there are (Martin and Priest 2017), which have to be updated as the software updates.

In a counter-terrorism context, consider a digital assistant that aids intelligence analysts in their search for terrorists. Analysts will use this assistant differently if a simple search could trigger an algorithmic evaluation, which automatically places someone on a no-fly list. Furthermore, if analysts don't have a clear picture of what the digital assistant accesses to perform its search and evaluation, then analysts will not have the information necessary to use their judgment. Rather than the algorithm aiding the analysts, the analysts are reduced to a rubber stamp role. As the capabilities of AI-powered tools increase, it will be important that users have the information they need to use these tools properly.

AI-powered machines can also have deadly results – as can be seen by the recent autonomous car crashes of Uber and Tesla. It is unknown what combination of inputs resulted in, for example, a Tesla slamming into the road barrier resulting in the passenger's death (Levin 2018). While we cannot control our environment (e.g., a drunk driver might slam into your car), AI-powered machines are the first example of us not being able to control the tools we use to navigate our environment. No utilitarian calculus can change the disturbing idea that your autonomous car, for reasons unknown, may slam into a pedestrian or barrier. Without basic knowledge surrounding these machines, how are users supposed to use them ethically? How are bystanders supposed to navigate a world filled with these machines appropriately? Finally, how are governments supposed to craft effective policies and regulations for these machines?

One major problem with focusing on explanation as a fix for the opaque inner workings of AI-powered machines is that many of these machines are beneficial *because* they

aren't articulable in human language. A cancer detection algorithm that cannot explain why one mole is labeled as cancerous should not be considered a problem if it is more effective than dermatologists.[74] Likewise, a detection algorithm used to detect chemical compounds used to make the bombs terrorists use. Since many of the benefits of AI-powered machines come from inherently opaque inner workings, we must zoom out as we did in this section to see the other opacities surrounding AI-powered machines. When seen from this perspective, a better solution to this problem is needed. The solution, I argue, can be found in a concept borrowed from the robotics field: envelopment.

6.3 Envelopment

Luciano Floridi has claimed that robots will be successful when "we envelop microenvironments around simple robots to fit and exploit at best their limited capacities and still deliver the desired output" (Floridi 2011a, p. 113). The term 'envelop' is borrowed from the field of robotics. The 'envelope' of a robot is the "three-dimensional space that defines the boundaries that a robot can reach" (Floridi 2011b, 228). Luciano Floridi has discussed envelopment as a process that allows for robots and AI to be more effective. He provides a striking example of washing dishes. Dishwashers are effective because they have been appropriately enveloped within an environment conducive to its operations (a closed box we call a dishwasher). The alternative is a humanoid robot, which would be decidedly ineffective at washing dishes.

Using Floridi's dishwashing robot as an example, we can see two broad sets of issues concerning non-enveloped robotics and AI. First, the humanoid robot would constantly face novel scenarios (i.e., its inputs are not

[74] Other ethical concerns may, however, be raised for this application, e.g. concerns regarding the appropriate training data when algorithms are proven to work far better on fair skin than on darker skin tones (Lashbrook 2018).

precisely defined and constrained) in which it would have to make judgments that could result in harm. I would consider myself deeply harmed were such a robot to scrub my new Le Creuset nonstick skillet with an abrasive brush. Add in mistaking a tablet computer for a plate, and we can see a few of the many complex decisions such a humanoid robot would encounter. Furthermore, this robot would have to share its environment with humans. This increases the potential for ethical dilemmas and harm to humans.

Second, the task of the robot is ill-defined. "Wash dishes" is not precise enough. This could mean finding dirty dishes throughout a household, washing and drying those dishes, and putting them away. Giving a robot this umbrella task, one could easily envision further tasks that would need to be added on: notifying a human that the soap is running out, sweeping broken glass, etc. Human users of such a robot may justifiably expect the robot to do things it is merely unable to do. These expectations could be mitigated if the robot's boundaries and functions were explicitly defined. These two sets of issues (harmful judgments and undefined task) should not occur in robotic and AI systems.

Floridi also proposes that driverless vehicles will only enjoy success if envelopment happens for them:

> *If drones or driverless vehicles can move around with decreasing troubles, this is not because productive AI has finally arrived, but because the "around" they need to negotiate has become increasingly suitable to reproductive AI and its limited capacities. (Floridi, 2011a, p. 228)*

The limits of driverless cars in a non-enveloped environment have been shown dramatically. In April 2018, one passenger and one pedestrian were killed by cars operated by artificial intelligence in separate incidents. To date, the focus on driverless cars has been to increase their 'intelligence' by self-learning algorithms and to

increase the effectiveness and capabilities of their sensors. The missing ingredient, according to Floridi, is envelopment.

In a counter-terrorism context, we can conceive of an AI-powered automatic weapons system that acts as a sentry. It is capable of seeking out and shooting people. One could attempt to make the system more and more intelligent to handle novel contexts without resulting in harm to civilians. However, given the outputs of this system (shooting people), it would be a lot easier to deploy it in contexts where there are no civilians (if that is indeed possible).

But envelopment with cars becomes increasingly tricky. First, we would need to make the roads and their surroundings machine-readable. Rather than relying on image recognition AI to 'see' that a stop sign is coming up, sensors could be built into the road which the car is easily able to read. This prevents a stop sign from being missed by the car's cameras due to a mud-splattered sign or heavy fog. The effectively enveloped environment for a driverless car would be one that closes out all unexpected variables. Pedestrians and cyclists would not be allowed on the road, all cars would be driverless (human drivers are unpredictable), and all of the road signs, dotted lines, solid lines, etc. would emit signals for the driverless cars to read. Truly enveloped driverless cars would not be able to leave the enveloped zone. This is because its inputs outside of an enveloped zone are potentially anything going on near automobile infrastructure. Ignorance about the possible inputs has led to fatal crashes. There is little advice given on where and when these cars should be used in autonomous mode. Are they only intended for recently built infrastructure on sunny, clear days? We don't know. Tesla does not claim what context autonomous cars should be used

in – they simply say that the human operator should have their hands on the wheel in case they need to take over.

Envelopment would solve a lot of problems; however, as Floridi notes, this raises the possibility that the world becomes a place that reduces our autonomy in that we will have created a world in which we are forced to adapt to the envelopment needed by machines. Floridi is concerned with ensuring that this process of envelopment occurs with our foresight and guidance to prevent a world that works well with robots and AI but is not desirable to human beings. People who cannot afford new driverless cars would be forced to cope with a crumbling infrastructure for their non-driverless cars as more and more resources are used for the infrastructure serving as the envelope for driverless cars. The privacy concerned may be put at risk because the world has been changed such that AI-driven machines rely on sensors implanted into human beings – sensors which the privacy-concerned refuse, rendering them invisible. Although I argue in this chapter that we should envelop AI-driven machines, it is essential to note that the envelopes themselves may be unethical. This is why we must know what the envelope would have to be before we thrust these machines into society. This knowledge gives us a chance to say that the required envelope wouldn't be worth it.

Envelopment also reduces the number of possibilities for AI-powered machines. Many contexts within counter-terrorism are not suitable for envelopment. Autonomous vehicles for soldiers looking for terrorists in an urban area, for example, could not rely on that area to be appropriately sensored for envelopment. This does not mean that envelopment is an untenable condition to be placed on AI-powered machines for counter-terrorism. Rather, it means that many AI-powered machines are not suitable for countering terrorism.

While Floridi uses envelopment to describe the conditions under which AI-powered machines will be successful, I argue that envelopment describes the conditions under which AI-powered machines should be considered acceptable. The example of autonomous weapons shows the potential harm which can occur when operating non-enveloped AI-powered machines. If autonomous weapons are placed in an environment where the possible inputs are infinite, then it will be challenging to prevent fatal mistakes. While we may not know how the algorithm results in a particular action, decision, or output, we should know enough about the possible inputs and outputs to know under what conditions a particular AI system should be used. Some basic knowledge about the machine helps us to make its envelope – preventing harm while helping the machine reach its full potential. If the envelope is too difficult to create (e.g., autonomous weapons), then the machine in question would be unethical to implement. To say otherwise is to say one of two things: either that we don't know enough about the machine to create such an envelope and therefore cannot prevent harmful situations (e.g., digital assistants), OR that we know that this machine will lead to harm in the context that it is placed in, but it is too costly or implausible to build the required envelope (e.g. autonomous cars). Both should be unacceptable to regulators, users, and bystanders.

To achieve the envelopment of an AI-powered machine requires a level of knowledge about the machine that we often lack. To be clear, knowledge alone does not prevent bad things from happening. Knowing that a machine is capable of an output that causes serious bodily harm should prevent us from putting it into contexts where that output would cause serious bodily harm. This is how knowledge is connected to solving the diverse ethical issues that will arise when using AI-powered machines. Before we have this knowledge, we will not know if regulation, policy, ethical norms, or an outright ban will

be the path to the responsible development, implementation, and use of AI-powered machines. Just like we don't put chainsaws into daycare centers, we should not put trash compacting robots in places where babies are sleeping. Nor should we put driverless cars in urban CT operations. This knowledge will allow us to envelope AI-powered machines. Only then can these machines be considered to be under meaningful human control (Santoni de Sio and van den Hoven 2018) – control that is needed to responsibly regulate, use, and be a bystander to such machines.

The knowledge that we are lacking not only refers to how the machine works, but the what, why, and where. The "what" referring to the training data, possible inputs, and possible outputs. The "why" referring to what the machine is intended to be used for, i.e., its function. And the "where" referring to what boundaries constrain this machine. There are simply too many unknowns concerning some AI-powered machines to regulate and use them. Many products now powered by AI are like the Monolith in Stanley Kubrick's *2001: A Space Oddysey* in that their purpose, capabilities, and inputs are a complete mystery. The point is that we are in no epistemic position to create ethical norms, enact policy and regulation, or engage with these AI-powered machines until we shine a light on these critical properties.

6.4 Towards the Envelopment of AI

If we are to make responsible decisions about regulating and using AI-powered machines, we need to know a lot more about them than we often do. This is especially true for modern AI algorithms (e.g., deep learning), which are opaque concerning their reasoning. The training data, inputs, outputs, functions, and boundaries of these machines must be known to us.

6.4.1 Training Data

The data used to train machine learning algorithms is extremely important with regard to how that algorithm or machine will work. Two algorithms that share the same code could work wildly differently because they were trained using different datasets. A facial recognition system trained only using pictures of faces of old white men will not work very well for young black women. If someone is to buy a facial recognition algorithm, then there should be some information about the faces used to train it. The number of faces and the breakdown of age, ethnicity, sex, etc. would be an essential start. The specifics regarding what information is needed about the training data will vary depending on context and type of data.

The knowledge regarding training data will be necessary when implementing algorithms. Simply knowing that the training data lacks a particular demographic would hopefully cause one to test the system before using it on such a demographic or to restrict its use to demographics covered by the training data. For example, if a previously successful algorithm used to detect terrorist phone activity is to be adopted by an intelligence agency, it might be useful to know that the algorithm was trained using data from Al Queda and Islamic State (IS). This would at least give the intelligence agency pause before using it to detect terrorist phone activity on far-right extremist groups.

Knowledge of training data can also help to determine unacceptable algorithms that will simply reinforce societal stereotypes (Koepke 2016; Ensign et al. 2017). Predictive policing algorithms that rely upon training data that is biased against African Americans simply should not be used. The knowledge of this bias would not lead to its envelopment; rather, it should, if possible, lead to fixing the training data.

6.4.2 Boundaries and Inputs

The terms 'boundaries' is construed broadly. Not only does it mean physical boundaries in the case of a robot, but also virtual boundaries which refer to the possible inputs (or types of input) in the form of data that it could encounter. 'Boundaries,' then, refers to an algorithm's or robot's expected scenarios. For example, AlphaGo expects as an input a GO board with a configuration of white and black pieces. AlphaGo is not expected to be able to suggest a chess move based on an input of a chessboard with a configuration of pawns, knights, bishops, rooks, queens, and kings on it. An algorithm playing chess is fine but is a different algorithm than AlphaGo.

Knowing precisely what the boundaries a machine is constrained by helps us know what the possible inputs are. For example, a money-laundering detection algorithm might only have access to the financial transactions occurring at a specific bank or category of transaction, e.g., those going in and out of organizations or to and from individuals with terrorist profiles. This makes it clear that the algorithm is classifying certain transactions as possibly associated with organized crime or terrorist groups due to the nature of the transactions themselves. This is opposed to an algorithm that not only uses lists of transactions but also surveillance video with sensors that claim to detect the emotional state of people entering the bank to make transactions and sentiment analysis on their social media posts and emails. The latter opens up the algorithm to so many possible inputs that it may be difficult for an individual to exercise control over the output of the algorithm.

Boundaries are different from inputs. A machine's inputs are determined by its sensors or code. The money laundering detection algorithm above may have cameras, microphones, and facial recognition, all serving as inputs into the machine. An 'input' as I want to talk about it

here, is the combined data from all sensors. We, as humans, make decisions based on several factors. For example, we might put on a rain jacket because it is raining, it is not too cold outside (otherwise we would opt for a heavy jacket), and we are going to be outside. A machine might be able to tell a user to wear a rain jacket based on the same data because it has a temperature sensor to sense how cold it is outside, a data feed received from a weather website (to 'sense' that it is raining), and a microphone to hear the user say they need to go outside. It is the combination of this data that determines what output will be given.

So we not only need to know what types of inputs there are (sound, image, temperature, specific voice commands, data feeds, etc.) but how these get combined to form one input. There are machines that take limited inputs, which make significant classifications. The weapon detection algorithm mentioned in the introduction only takes x-ray scans of baggage. We have a very clear understanding of the inputs of this machine. On the other hand, a driverless car has many sensors that combine to provide infinite combinations of inputs.

I don't mean to suggest that a machine that can accept infinite combinations of inputs should not be used. We simply must know that this is the situation. We may know that an AI app on our phone accepts data from weather stations, our voice commands, images of our face, etc., as well as feedback after its decision (so that it can improve). Furthermore, it may not have any real boundaries – that is, it can grab data from other sources if it helps to improve its decisions. However, the function of the machine may simply be to decide whether or not to advise the user to wear a jacket. That is, it only has two outputs: jacket or no jacket. We can debate about the overkill regarding using AI for advice on our outdoor clothing; however, the point is that a decision about the

acceptability of a machine requires not only knowing its boundaries and inputs but its function and outputs as well.

6.4.3 Function & Outputs

Knowledge of the functions and possible outputs of a machine is essential if we are to achieve the goal of enveloping AI-powered machines. In the AlphaGo example, the output is a legal move in the game of GO. We might be shocked by it making a particular move, but it is nonetheless a legal move in the game of GO. It would be strange if the function of AlphaGO was defined as "not letting an opposing player win." Instead of making a move, its output was to mess up the board (because it knew there was no chance of winning, and this was the only way to ensure that the other player did not win). In the autonomous weapons debate, it would matter significantly whether the function of the algorithm was "detect terrorists" or "detect and kill terrorists."

It can be easy to think that functions and outputs are equivalent. In the case of the jacket deciding machine in the previous section, the function of the machine is to advise the user on whether or not to wear a jacket. This is the same as its output, which is either "jacket" or "no jacket." This, however, is often not the case. The function of a driverless car is to drive from point A to point B; however, this will involve many outputs. Each turn, acceleration, swerve, and brake is an output. Defined functions are of the utmost importance because they allow us to test the machines for efficacy. How well a machine functions is salient with regard to its moral acceptability. If the weapon detecting algorithm were seldom successful at categorizing baggage as having weapons or not, then it would be unethical to use it. Equally unethical is the use of the algorithm when we are ignorant about how successful it is (i.e., use outside of a testing environment).

Outputs are not the same as a machine's function; however, they can be discussed in the same way that we talk about a machine's capabilities. What can the machine do? A driverless car may be able to go 200 mph – which means that this is a possible output. A drone may have a machine gun built-in, giving it the capability to shoot bullets – which means a possible output is the shooting of bullets. This example makes it clear why it is so important to know the functions, outputs, boundaries, and inputs. A machine whose possible output is to shoot bullets may be acceptable if its only input is a user telling it to shoot and its boundaries are a bulletproof room. We need all of this knowledge to make informed decisions regarding the acceptability of machines.

6.4.4 Stepping out of the Dark

Knowing what the inputs, boundaries, training data, outputs, and functions of an AI-powered machine will allow us to have some clue as to the envelopes needed for these machines to operate appropriately. Even when machines are operated in environments that are so broad that we cannot prevent novel scenarios, the knowledge that this is the case helps inform our decisions regarding such a machine's acceptability. If there are possible novel environments (and therefore we are ignorant of the possible inputs), then the outputs must be such that it does not matter. No matter what novel board configuration of the game GO is given to AlphaGo, the output is always a legal move of GO. It is simply not possible for a harmful output. It would not matter if AlphaGo took as its inputs live CCTV video feeds from all over the world – the outputs would always be the same benign GO moves (although such inputs would probably not help with the stated goal of winning the game of GO). This is in direct contrast to the situation we face with driverless cars. Their possible inputs are states of affairs on just about any road in the world – with the weather, pedestrians, other cars, etc., all

combining to create consistently novel inputs. In this case, though, the outputs are potentially fatal.

Machines that have precise specifications regarding the properties listed in sections 4.1-4.3 limit these problems. Cortis is an algorithm that detects voice patterns associated with cardiac arrest (Vincent 2018). The algorithm exists explicitly to aid emergency call operators (we know its function). The algorithm takes as its input live sound from the calling line. Its output is true if the voice pattern is associated with cardiac arrest and false if it is not (explicitly defined outputs). This algorithm being so explicit means that we know enough to determine that this is an acceptable machine. If the machine is used within the boundaries given, then we can easily figure out what the possible scenarios are – without understanding how the machine comes to its decision. The machine either outputs true or false. If true, and a person at the end of the phone line indeed has a heart attack, then the machine may be instrumental in preventing death. If the output is true, and no one on the end of the phone line is having a heart attack, then emergency services may be sent out without it being necessary. While this is not an ideal situation, knowing that it could occur gives us the knowledge to decide whether this risk is worth it. If the machine outputs false, and no one at the end of the line is having a heart attack, then the emergency call is unaffected by the machine. The last scenario is the machine outputting 'false' when someone on the line is having a heart attack. This is the worst scenario; however, the consequences of the machine acting this way are no different from the consequences of the emergency call without the machine. Again, the knowledge that this could happen is necessary for us to decide whether this is an acceptable risk.

We can imagine a machine that would operate in a context that could result in unacceptable risk – because we don't

have the necessary information to make an informed choice. The machine would be an algorithm that goes through social media posts to determine who is at most risk of becoming radicalized. The sheer number of possible inputs to this machine makes it difficult to determine how it could harm someone. In one obvious way, the machine could overestimate the seriousness of someone's propensity for radicalization – resulting in a series of intrusions into that person's life that were unnecessary. It may be the case that the algorithm results in fewer false-positives than when human analysts perform the same task. However, empirically validating this is next to impossible – especially before these algorithms are implemented.

If we are in the dark about the inputs, boundaries, functions, and outputs, then we have a machine we don't know enough about to properly envelop – leading to its possible failure, which will often be an unacceptable risk to human beings. For, with modern AI, we are already in the dark about how it makes decisions. An undeveloped machine means that we are also in the dark about what could happen with these machines.

Ideally, AI-powered machines will be designed for envelopment – with clear ideas about the training data, inputs, functions, outputs, and boundaries. This knowledge would be necessary to design for values properly or to facilitate an ethicist as part of the design team (van Wynsberghe and Robbins 2014). Not only would this result in ethically better designs but may prevent a waste of resources on a machine that cannot be enveloped and, therefore, may be designed to fail.

6.5 Objections

One objection could be that envelopment prevents the ultimate dream (or a nightmare depending upon your perspective) of AI: developing general artificial intelligence. While some still dream of general artificial

intelligence, which will outperform humans at just about any task (Bostrom 1998; Müller and Bostrom 2016), the knowledge I am arguing for would explicitly exclude such a machine. General is the opposite of specific, and general AI would be expected to perform many different tasks, have a variety of outputs, and accept unlimited inputs. Luckily, this is not even on the horizon for robotics and AI right now, despite some futurists making bombastic and outlandish claims about this possibility. As Floridi puts it:

> *True AI is not logically impossible, but it is utterly implausible. We have no idea how we might begin to engineer it, not least because we have very little understanding of how our own brains and intelligence work. This means that we should not lose sleep over the possible appearance of some ultraintelligence. (Floridi 2016)*

We should not be basing our ethical considerations and discussions around the possibility of general or strong AI. The focus should be on what is happening now and what could be happening in the foreseeable future. We must remember that robots just recently learned how to open a door – a capability that may be dependent upon specific door handles (Sulleyman 2018). We must not put a cart of ethical issues before the horse of the possibility of strong AI. It would be absurd to discuss the ethics surrounding eating unicorn meat when the foreseeable future does not include unicorns. The point is that discussion of general, super, or strong AI is a distraction from the real problems surrounding AI and robotics.

More pressing is the objection that requiring such knowledge would stifle innovation in AI. When Elon Musk claims that those opposing autonomous cars are "killing people" (McGoogan 2016), he is claiming that innovation in autonomous cars will save lives in the long run – so we should do it despite concerns. Envelopment would, to

be sure, stop his Teslas from having the "autopilot" option. This function is not enveloped – and therefore, we do not know enough to make informed choices regarding its implementation and usage. However, envelopment would leave plenty of room for artificial intelligence to thrive.

The AI machines which are successful are the ones that are already enveloped. The weapon detecting algorithm, AlphaGo, machines for analyzing x-rays (Litjens et al. 2017), spam filtering, fraud detection, etc. are all enveloped – and many of them are valuable with regard to helping us solve serious problems. Furthermore, we can measure how effective all of these machines are. Most importantly, envelopment is a workaround for AI's transparency problem. If enveloped, AI machines can remain black boxes – therefore ensuring that the benefits of AI are kept.

Instead of fully autonomous weapons, which would be impossible to develop, there are many algorithms that could aid in finding targets for strikes. An algorithm that is solely tasked with finding groups of people with weapons could aid significantly in finding worthy targets. Because the input would be images, humans could verify that the algorithm was correct that the people depicted were armed. When thinking in terms of envelopment, we can divide the tasks between human and machine more effectively – that is, choose tasks that make sense for an algorithm to have.

One objection which is difficult to resolve is that contemporary AI machines often have multiple algorithms at work to take inputs and create outputs. Just what is it that should be enveloped? That is, what is the machine? In a driverless car, many sensors are feeding into many algorithms, which in turn feed their outputs to an algorithm that results in action. Taken as one machine, we might reach one evaluation; namely, that we lack the

knowledge we need to envelop the machine. However, if we take this machine apart, we may have many machines which are enveloped. For example, if there was an algorithm that takes as its input an image of the inside of the car while it is in motion which outputs how many people are in the car, the algorithm itself doesn't seem to have much problem. There are clear inputs, outputs, boundaries, and a function.

However, human users of a driverless car experience the outputs of the car – the turns, accelerations, braking, etc. Human users may not even be aware of the camera on the inside of the car – or the sensors detecting the outside world. The outputs of concern are the turns and accelerations of the car – not of the individual sensors. So the driverless car as a whole should be the object of evaluation.

Importantly, however, each of the machines which makes up the driverless car should be enveloped as well. What is different is the users of the machine. In this case, the user of the machine is the automaker. Each AI machine which makes up the driverless car should be enveloped – that is, we should know their possible inputs, possible outputs, boundaries, and function. Not knowing these things about a machine of importance to the functioning of the driverless car would be unacceptable.

6.6 The Limits of Envelopment

It must be said that envelopment is not enough on its own. Although the function of the machine must be known to us, this chapter says nothing of what functions should be assigned to robotics and AI Systems. It is easy to conceive of a robotic or AI system in which we have the knowledge I have argued we should require but is tasked with creating a superbug or killing someone. What functions should be excluded from acceptable applications of AI is an important question. This question is actively debated in

the field of robot ethics and the ethics of AI. Tasks that are deemed unethical for AI systems, therefore, should not be considered by developers, and attempting to envelop machines is a step that only applies to those machines whose functions are deemed ethical. Knowledge about these machines can help us with this, however. If the boundaries and function of the machine are forced to be made explicit, then it will be much easier to focus on whether or not this machine's function and context are acceptable.

The envelopment of a machine does not mean that a particular machine is effective. An enveloped machine may be spectacularly bad at achieving its function. This should be a reason not to use a particular machine. What knowledge of the features described above can do for us concerning efficacy is to help us understand what success means for a particular machine. How do we judge the success of a machine when we do not know what its function is or the boundaries of its operation? A machine that is precise about its inputs, outputs, boundaries, function, and training data comes ready-made with a rubric for the evaluation of its efficacy.

A more general issue that this knowledge and ideal state of envelopment does not cover is the subtle changes technology can have on society. Just because we have the necessary knowledge for envelopment does not ensure that society will be changed for the better due to the technology. Guns serve as a good example here. We have good knowledge about how they work – their inputs and outputs. We can even say that there is meaningful human control in relation to guns. However, the option of using a gun opens up choices that weren't available before. The ability to quickly and easily kill people has led, despite meaningful human control, to situations like the US, where too many people are harmed and killed. It would be better if such choices did not exist – and many countries have passed legislation taking away this choice. Similar

arguments have been made against human-controlled drones and they're lowering the barrier to war and killing (see, e.g., Boyle 2015).

The same will need to happen concerning specific machines. Already there are calls to enact a ban on autonomous weapons (Sampler 2017). There may be many other machines that are unacceptable for their societal impact despite meaningful human control. Evaluations on societal impact – given envelopment and efficacy – will be critical. I do not pretend that the arguments in this chapter help with such an evaluation; rather, they can prevent us from wasting time evaluating machines that have a more immediate problem: we don't have the knowledge to make informed evaluations in the first place.

6.7 Conclusion

The techno-optimism surrounding AI is running high. There seems to be no limit to its applications and no bounds to the hype in the media. It can be difficult, therefore, to separate real hope from fantasy, the good ideas from the ridiculous, and the responsible from the irresponsible. Luciano Floridi has helpfully highlighted the concept of envelopment to help us to understand what makes for successful robotics – and, as I have argued – for responsible robotics. To get to an enveloped state, however, we must know some basics concerning these machines: the inputs, functions, training data, outputs, and boundaries.

Not only would such knowledge inform further ethical evaluation about whether or not a specific function is an acceptable task for a machine, but it achieves a necessary condition for meaningful human control. Despite concerns about stifling innovation, envelopment allows for opaque algorithms to do what they do best. It simply keeps that opacity constrained to how the machine makes decisions. I argued here that opacity, which spreads well beyond the

'how' of the machine and into the what, where, why, etc. is unacceptable. This allows us to realize the great things AI promises to us while keeping the fantastical, unnecessary, and dangerous machines out.

Envelopment is simply one part of the puzzle which, when solved, will result in creating AI-driven machines that will benefit and not harm society. Given envelopment, there are still important ethical evaluations that need to be made regarding the appropriateness of delegating a particular task to a machine, whether or not the operation of that machine is under meaningful human control, and what subtle societal effects such machines will have. While envelopment won't answer these important questions, it is a necessary and crucial first step towards the responsible design, development, and implementation of AI-powered machines.

Conclusion

Considering the increase in the use of ML in counter-terrorism and the need for legitimacy of the state's use of this technology, this thesis centered on the concept of MHC as a way forward in the responsible development and use of ML. To do this, the focus of this thesis was on the question: *how can we ensure meaningful human control over artificial intelligence in a counter-terrorism context?* In addressing this question, chapter 1 pointed out that the current propositions for MHC (being either technology-centered or human-centered) miss an essential first step. Before concerning ourselves with MHC as it has been described in the literature to date, we have to ensure that the data is responsibly sourced, collected, and labeled. Any gains we get from AI resulting from incorrect or biased data is the fruit of a poisonous tree.

Furthermore, the algorithm has to be able to perform with some level of efficacy. An ineffective algorithm will be unethical to use no matter the level of control humans have over it. We may go so far as to suggest that these conditions (i.e., the quality of data and the efficacy of the algorithm) be considered necessary conditions for ensuring MHC. In other words, if you don't have control

over the data, then you don't have control over the algorithm, and consequently, you do not have MHC.

In line with the concerns over data quality, I also discussed issues related to data acquisition and sourcing. Chapter 2 argued for specific constraints to be placed on the state concerning its bulk collection of data – whether used for training and analysis by AI or otherwise. These constraints included: not using AI as a filter; not allowing consumer companies like Google and Facebook to act as intelligence agencies (i.e., to collect data otherwise not collected for counter-terrorism); collected data must be tied to a filter and deleted when the justification for that filter no longer holds, and; data must only be used for the legitimate purpose for which it was initially collected. These constraints will prevent situations in which we have powerfully trained AI that works well but was trained on data that was collected in a way that violates liberal democratic values. We don't accept medical information that was gained by the forced testing of vulnerable groups of people – and we shouldn't accept algorithms that were dependent upon the unethical collection of data.

Chapter 3 shows that transparency is a necessary condition for the government and its security agencies to use algorithms for counter-terrorism purposes under liberal democratic values. It is, in part, the people that need to have control over the government and its security. And it is, in part, the government that requires the consent of its citizens to legitimize the methods used to secure the citizenry and the state itself. Therefore, the government, by way of independent agencies, should audit data collection to identify incidentally collected information – that is, information that should not be collected in the first place. The government should also release details regarding a method's effectiveness in preventing or combating terrorism.

Chapters 4 and 5 show that neither endowing machines with ethical reasoning nor making algorithms explicable are sufficient to establish MHC. Regarding the first point, scholars have proposed endowing machines with artificial moral reasoning capabilities to prevent machines from performing unethical actions. I have argued that there is no good reason for doing this in addition to the implausibility of validating that the machine was working correctly – i.e., generating outputs that are 'ethical.' Regarding the second point – that explicable AI is not sufficient for MHC - I evaluated the oft proposed principle of explicability for AI to show that explicability does not solve the moral problem of MHC for ML. If a human is supposed to check the explanation provided by the algorithm to see if it is ethical, then that human already needs to know what a 'good' explanation is for the particular output. But if we already have that information, then we do not need ML. So if an autonomous weapon were to be able to explain why it is about to kill a terrorist suspect, then a human would need to check whether or not the reasons given were acceptable. But if we already know these reasons, then we can simply use an automated (but not fully autonomous) weapon with a human in control to follow the pre-determined, morally justified procedure (e.g., GOFAI) for the targeted killing of terrorists.

Finally, in chapter 6, I propose a principle of envelopment. That is, we must create an envelope for AI systems to operate in. The designed envelope will constrain the AI system in such a way as to minimize undesirable potential harms, e.g., intrusive surveillance of innocent persons in the course of using AI-powered surveillance technology for counter-terrorism purposes. For us to properly envelop AI systems, we have to know about the training data, possible inputs, possible outputs, functions, and boundaries of these systems. This information may lead us to determine, as in the case of

autonomous cars and weapons, that an envelope cannot be reasonably created. However, if we cannot create an envelope for the machine, then AI is not an appropriate solution.

So *how can we ensure meaningful human control over artificial intelligence in a counter-terrorism context?* We ensure that: the algorithm is trained using data collected in compliance with liberal democratic values; the tasks delegated to AI are in part legitimized by the people they are supposed to protect, through active transparency; we do not delegate outputs to AI which require explanations; and that we have created an envelope for the algorithm.

These conditions might appear to be unduly restrictive for AI. The goal of Artificial General Intelligence is not possible given my constraints; it might be claimed. However, if my argument in this thesis is sound, then AGI should be considered a non-starter. The future of AI is not, and should not be, machines from which human moral responsibility has been removed, but in machines that enhance our ability to be morally responsible.

What is left out of the answer to my research question is what, exactly, should and should not be delegated to machines? I have described cases of technologies useable for counter-terrorism purposes – such as targeting and killing of human beings by autonomous weapons– as examples of what should *not* be delegated to machines. However, future research needs to be more precise about what kinds of tasks are suitable for machines. We know from chapter 5 that AI-powered machines should not be delegated many tasks that generate outputs that require explanations – however that simply kicks the can down the road as we will now need an answer to the question: *what kinds of outputs require explanations?*

I hypothesize that the kinds of outputs that require explanations are evaluative ones that have morally significant consequences, e.g., have potentially harmful consequences if realized - such as surveilling, detaining, or shooting dead persons suspected of being terrorists who might not be terrorists. I suggest that part of keeping MHC over machines means restricting machines to outputs to do not amount to value judgments. Machines that can make decisions based on opaque considerations should not be telling humans what decisions morally ought to be made and, therefore, how the world *morally ought* to be. Delegating these judgments of moral value to machines is a reduction of human control over our most important sphere of decision-making. Not only would we be losing control over specific decisions in specific contexts, but we would be losing control over moral decision-making in particular. Outsourcing moral decision-making in general, and certainly in counter-terrorism contexts, to AI-powered machines, will forever result in a loss of MHC.

References

Adadi, Amina, and Mohammed Berrada. 2018. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)." *IEEE Access* 6: 52138–60. https://doi.org/10.1109/ACCESS.2018.2870052.

Ahmed, Maha. 2018. "Aided by Palantir, the LAPD Uses Predictive Policing to Monitor Specific People and Neighborhoods." The Intercept. May 11, 2018. https://theintercept.com/2018/05/11/predictive-policing-surveillance-los-angeles/.

"AI at Google: Our Principles." 2018. Google. June 7, 2018. https://www.blog.google/technology/ai/ai-principles/.

"AI Principles." 2017. Future of Life Institute. 2017. https://futureoflife.org/ai-principles/.

Allen, C., W. Wallach, and I. Smit. 2006. "Why Machine Ethics?" *IEEE Intelligent Systems* 21 (4): 12–17. https://doi.org/10.1109/MIS.2006.83.

Allen, Colin, Iva Smit, and Wendell Wallach. 2005. "Artificial Morality: Top-down, Bottom-up, and Hybrid Approaches." *Ethics and Information Technology* 7 (3): 149–55. https://doi.org/10.1007/s10676-006-0004-4.

Allen, Colin, Gary Varner, and Jason Zinser. 2000. "Prolegomena to Any Future Artificial Moral Agent." *Journal of Experimental & Theoretical Artificial Intelligence* 12 (3): 251–61. https://doi.org/10.1080/09528130050111428.

Allen, Colin, and Wendell Wallach. 2014. "Moral Machines: Contradiction in Terms or Abdication of Human Responsibility?" In *Robot Ethics: The Ethical and*

Social Implications of Robotics, edited by Patrick Lin, Keith Abney, and George A. Bekey, 55–66. Cambridge, Mass.: The MIT Press.

Altman, Andrew, and Christopher Heath Wellman. 2009. *A Liberal Theory of International Justice*. 1 edition. Oxford ; New York: Oxford University Press.

Anderson, David. 2016. *REPORT OF THE BULK POWERS REVIEW*. United Kingdom: Williams Lea Group. https://terrorismlegislationreviewer.independent.gov.uk/wp-content/uploads/2016/08/Bulk-Powers-Review-final-report.pdf.

Anderson, Kenneth, and Matthew C. Waxman. 2012. "Law and Ethics for Robot Soldiers." SSRN Scholarly Paper. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=2046375.

Anderson, Michael, and Susan Leigh Anderson. 2007. "Machine Ethics: Creating an Ethical Intelligent Agent." *AI Magazine* 28 (4): 15–26.

———. 2010. "ROBOT BE GOOD." *Scientific American* 303 (4): 72–77.

———. 2011. *Machine Ethics*. Cambridge University Press.

Anderson, Susan Leigh. 2011. "Machine Metaethics." In *Machine Ethics*. Cambridge University Press.

Angwin, Julia, and Jeff Larson. 2016. "Machine Bias." Text/html. ProPublica. May 23, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Ardabili, Sina F., Amir Mosavi, Pedram Ghamisi, Filip Ferdinand, Annamaria R. Varkonyi-Koczy, Uwe Reuter, Timon Rabczuk, and Peter M. Atkinson. 2020. "COVID-19 Outbreak Prediction with Machine Learning." SSRN Scholarly Paper ID 3580188. Rochester, NY: Social Science Research Network. https://doi.org/10.2139/ssrn.3580188.

Aristotle, W.D. Ross, J.L. Ackrill, and J.O. Urmson. 1998. *The Nicomachean Ethics*. Oxford University Press. http://books.google.nl/books?id=Dk2VFlZyiJQC.

Arkin, Ronald. 2008. "Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture." In *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction*, 121–128. HRI '08. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1349822.1349839.

———. 2009. *Governing Lethal Behavior in Autonomous Robots*. CRC Press.

Arkin, Ronald, P. Ulam, and A. R. Wagner. 2012. "Moral Decision Making in Autonomous Systems: Enforcement,

Moral Emotions, Dignity, Trust, and Deception." *Proceedings of the IEEE* 100 (3): 571–89. https://doi.org/10.1109/JPROC.2011.2173265.

"Article 36." 2015. Killing by Machine: Key Issues for Understanding Meaningful Human Control. April 2015. http://www.article36.org/autonomous-weapons/killing-by-machine-key-issues-for-understanding-meaningful-human-control/.

Asimov, Isaac. 1963. *I, Robot*. Doubleday.

Aviv, Juval, and Juval Aviv. 2009. "Can AI Fight Terrorism?" June 2009. https://www.forbes.com/2009/06/18/ai-terrorism-interfor-opinions-contributors-artificial-intelligence-09-juval-aviv.html.

Azizan, Sofea Azrina, and Izatdin Abdul Aziz Aziz. 2017. "Terrorism Detection Based on Sentiment Analysis Using Machine Learning." *Journal of Engineering and Applied Sciences* 12 (3): 691–98. https://doi.org/10.3923/jeasci.2017.691.698.

Baier, Annette. 1986. "Trust and Antitrust." *Ethics* 96 (2): 231–60. https://doi.org/10.1086/292745.

Barry-Jester, Anna Maria, Ben Casselman, and Dana Goldstein. 2015. "The New Science of Sentencing." The Marshall Project. August 4, 2015. https://www.themarshallproject.org/2015/08/04/the-new-science-of-sentencing.

BBC News. 2015. "Google Apologises for Photos App's Racist Blunder," July 2015.

Bellaby, Ross. 2012. "What's the Harm? The Ethics of Intelligence Collection." *Intelligence and National Security* 27 (1): 93–117. https://doi.org/10.1080/02684527.2012.621600.

Bellaby, Ross W. 2016. "Justifying Cyber-Intelligence?" *Journal of Military Ethics* 15 (4): 299–319. https://doi.org/10.1080/15027570.2017.1284463.

Belson, Ken. 2006. "Senator's Slip of the Tongue Keeps on Truckin' Over the Web." *The New York Times*, July 2006.

Bergen, Peter. 2013. "Opinion: Would NSA Surveillance Have Stopped 9/11 Plot? - CNN.Com." 2013. http://www.cnn.com/2013/12/30/opinion/bergen-nsa-surveillance-september-11/index.html.

Berk, Richard A., Susan B. Sorenson, and Geoffrey Barnes. 2016. "Forecasting Domestic Violence: A Machine Learning Approach to Help Inform Arraignment Decisions." *Journal of Empirical Legal Studies* 13 (1): 94–115. https://doi.org/10.1111/jels.12098.

Blair, Irene V., Charles M. Judd, and Kristine M. Chapleau. 2004. "The Influence of Afrocentric Facial Features in Criminal Sentencing." *Psychological Science* 15 (10): 674–79. https://doi.org/10.1111/j.0956-7976.2004.00739.x.

Blum, Stephanie Cooper. 2008. "What Really Is at Stake with the FISA Amendments Act of 2008 and Ideas for Future Surveillance Reform." *Boston University Public Interest Law Journal* 18: 269.

Bostrom, Nick. 1998. "How Long Before Superintelligence?" *International Journal of Futures Studies* 2.

Boyle, Michael J. 2015. "The Legal and Ethical Implications of Drone Warfare." *The International Journal of Human Rights* 19 (2): 105–26. https://doi.org/10.1080/13642987.2014.991210.

Bridle, James. 2019. *New Dark Age: Technology and the End of the Future*. Reprint edition. Verso.

Bruijne, Marleen de. 2016. "Machine Learning Approaches in Medical Image Analysis: From Detection to Diagnosis." *Medical Image Analysis*, 20th anniversary of the Medical Image Analysis journal (MedIA), 33 (October): 94–97. https://doi.org/10.1016/j.media.2016.06.032.

Bryson, Joanna. 2010a. "Robots Should Be Slaves." In *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues*, edited by Yorick Wilks, 63–74. Amsterdam: John Benjamins Publishing.

———. 2010b. "Robots Should Be Slaves." In , edited by Yorick Wilks, 63–74. Amsterdam: John Benjamins Publishing.

Buolamwini, Joy, and Timnit Gebru. 2018. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." In *Conference on Fairness, Accountability and Transparency*, 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html.

Burrell, Jenna. 2016. "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms." *Big Data & Society* 3 (1): 2053951715622512. https://doi.org/10.1177/2053951715622512.

Cellan-Jones, Rory. 2014. "Stephen Hawking Warns Artificial Intelligence Could End Mankind." *BBC News*, December 2, 2014, sec. Technology. http://www.bbc.com/news/technology-30290540.

Chengeta, Thompson. 2016. "Defining the Emerging Notion of Meaningful Human Control in Weapon Systems." *New*

York University Journal of International Law and Politics 49: 833.

Chokshi, Niraj. 2018. "Is Alexa Listening? Amazon Echo Sent Out Recording of Couple's Conversation." The New York Times. May 26, 2018. https://www.nytimes.com/2018/05/25/business/amazon-alexa-conversation-shared-echo.html.

Citron, Danielle Keats, and Frank A. Pasquale. 2014. "The Scored Society: Due Process for Automated Predictions." *Washington Law Review* 89: 1.

Coeckelbergh, Mark. 2010. "Robot Rights? Towards a Social-Relational Justification of Moral Consideration." *Ethics and Information Technology* 12 (3): 209–21. https://doi.org/10.1007/s10676-010-9235-5.

Condon, Stephanie. 2018. "How Police Are Using Voice Recognition to Make Their Jobs Safer." ZDNet. July 20, 2018. https://www.zdnet.com/article/how-police-are-using-voice-recognition-to-make-their-jobs-safer/.

Council, National Research. 2015. *Bulk Collection of Signals Intelligence: Technical Options*. https://www.nap.edu/catalog/19414/bulk-collection-of-signals-intelligence-technical-options.

Crootof, Rebecca. 2016. "A Meaningful Floor for Meaningful Human Control." *Temple International & Comparative Law Journal* 30: 53.

Cummings, Mary. 2012. "Automation Bias in Intelligent Time Critical Decision Support Systems." In *AIAA 1st Intelligent Systems Technical Conference*. American Institute of Aeronautics and Astronautics. https://doi.org/10.2514/6.2004-6313.

Cuthbertson, Anthony. 2018. "Amazon Ordered to Give Alexa Evidence in Double Murder Case." The Independent. November 14, 2018. https://www.independent.co.uk/life-style/gadgets-and-tech/news/amazon-echo-alexa-evidence-murder-case-a8633551.html.

Dancy, Jonathan. 2004. *Ethics Without Principles*. Clarendon Press.

Darling, Kate. 2012. "Extending Legal Protection to Social Robots." IEEE Spectrum: Technology, Engineering, and Science News. September 10, 2012. http://spectrum.ieee.org/automaton/robotics/artificial-intelligence/extending-legal-protection-to-social-robots.

Darwall, Stephen. 2003. "Desires, Reasons, and Causes." *Philosophy and Phenomenological Research* 67 (2):

436–43. https://doi.org/10.1111/j.1933-1592.2003.tb00300.x.
Deng, Boer. 2015. "Machine Ethics: The Robot's Dilemma." *Nature News* 523 (7558): 24. https://doi.org/10.1038/523024a.
Department of Defense. 2018. "SUMMARY OF THE 2018 DEPARTMENT OF DEFENSE ARTIFICIAL INTELLIGENCE STRATEGY." https://media.defense.gov/2019/Feb/12/2002088963/-1/-1/1/SUMMARY-OF-DOD-AI-STRATEGY.PDF.
Desmarais, Bruce A., and Skyler J. Cranmer. 2013. "Forecasting the Locational Dynamics of Transnational Terrorism: A Network Analytic Approach." *Security Informatics* 2 (1): 8. https://doi.org/10.1186/2190-8532-2-8.
Detrixhe, John. 2018. "Australia Is a Battleground for Encrypted Apps." Quartz. 2018. https://qz.com/1497092/the-signal-encrypted-app-service-wont-comply-with-australias-assistance-and-access-bill/.
Dhar, Joydip, and Ashok Ranganathan. 2015. "Machine Learning Capabilities in Medical Diagnosis Applications: Computational Results for Hepatitis Disease." *International Journal of Biomedical Engineering and Technology* 17 (4): 330–40. https://doi.org/10.1504/IJBET.2015.069398.
Dietrich, Eric. 2001. "Homo Sapiens 2.0: Why We Should Build the Better Robots of Our Nature." *Journal of Experimental & Theoretical Artificial Intelligence* 13 (4): 323–28. https://doi.org/10.1080/09528130110100289.
Doris, John M. 1998. "Persons, Situations, and Virtue Ethics." *Nous* 32 (4): 504–30.
Edwards, Lilian, and Michael Veale. 2017. "Slave to the Algorithm: Why a Right to an Explanation Is Probably Not the Remedy You Are Looking For." *Duke Law & Technology Review* 16: 18.
Ekelhof, Merel. 2019. "Moving Beyond Semantics on Autonomous Weapons: Meaningful Human Control in Operation." Global Policy. March 19, 2019. https://doi.org/10.1111/1758-5899.12665.
Ellsberg, Daniel. 2013. "Edward Snowden: Saving Us From The United Stasi Of America." *The Guardian*, 2013.
Ensign, Danielle, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. 2017. "Runaway Feedback Loops in Predictive Policing." *ArXiv:1706.09847 [Cs, Stat]*, June. http://arxiv.org/abs/1706.09847.

Erickson, Bradley J., Panagiotis Korfiatis, Zeynettin Akkus, and Timothy L. Kline. 2017. "Machine Learning for Medical Imaging." *RadioGraphics* 37 (2): 505–15. https://doi.org/10.1148/rg.2017160130.

Esteva, Andre, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. 2017. "Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks." *Nature* 542 (7639): 115–18. https://doi.org/10.1038/nature21056.

Evans, Jake. 2019. "Australian Defence Force Invests $5 Million in 'killer Robots' Research." Text. ABC News. March 1, 2019. https://www.abc.net.au/news/2019-03-01/defence-force-invests-in-killer-artificial-intelligence/10859398.

Finlay, Stephen. 2007. "Four Faces of Moral Realism." *Philosophy Compass* 2 (6): 820–849. https://doi.org/10.1111/j.1747-9991.2007.00100.x.

Fjeld, Jessica, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. 2020. "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI." SSRN Scholarly Paper ID 3518482. Rochester, NY: Social Science Research Network. https://doi.org/10.2139/ssrn.3518482.

Floridi, Luciano. 2011a. "Enveloping the World: The Constraining Success of Smart Technologies." In *CEPE 2011: Ethics in Interdisciplinary and Intercultural Relations*, edited by J Mauger, 111–16. Milwaukee, Wisconsin.

———. 2011b. "Children of the Fourth Revolution." *Philosophy & Technology* 24 (3): 227–32. https://doi.org/10.1007/s13347-011-0042-7.

———. 2016. "True AI Is Both Logically Possible and Utterly Implausible – Luciano Floridi | Aeon Essays." Aeon. May 9, 2016. https://aeon.co/essays/true-ai-is-both-logically-possible-and-utterly-implausible.

Floridi, Luciano, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, et al. 2018. "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations." *Minds and Machines* 28 (4): 689–707. https://doi.org/10.1007/s11023-018-9482-5.

Floridi, Luciano, and J.W. Sanders. 2004. "On the Morality of Artificial Agents." *Minds and Machines* 14 (3): 349–79.

https://doi.org/10.1023/B:MIND.0000035461.63578.9d.

"Four Steps We're Taking Today to Fight Terrorism Online." 2017. Google. June 18, 2017. https://www.blog.google/around-the-globe/google-europe/four-steps-were-taking-today-fight-online-terror/.

France-Presse, Agence. 2017. "World's Best Go Player Flummoxed by Google's 'Godlike' AlphaGo AI." *The Guardian*, May 23, 2017, sec. Technology. https://www.theguardian.com/technology/2017/may/23/alphago-google-ai-beats-ke-jie-china-go.

Frenkel, Sheera. 2017. "Facebook Will Use Artificial Intelligence to Find Extremist Posts." *The New York Times*, June 2017.

Friedman, Batya, and Helen Nissenbaum. 1996. "Bias in Computer Systems." *ACM Trans. Inf. Syst.* 14 (3): 330–347. https://doi.org/10.1145/230538.230561.

Gaggioli, Andrea. 2017. "Artificial Intelligence: The Future of Cybertherapy?" *Cyberpsychology, Behavior, and Social Networking* 20 (6): 402–3. https://doi.org/10.1089/cyber.2017.29075.csi.

Ganor, Boaz. 2019. "Artificial or Human: A New Era of Counterterrorism Intelligence?" *Studies in Conflict & Terrorism* 0 (0): 1–20. https://doi.org/10.1080/1057610X.2019.1568815.

Gao, George. 2015a. "What Americans Think about NSA Surveillance, National Security and Privacy."

———. 2015b. "What Americans Think about NSA Surveillance, National Security and Privacy." *Pew Research Center* (blog). May 29, 2015. http://www.pewresearch.org/fact-tank/2015/05/29/what-americans-think-about-nsa-surveillance-national-security-and-privacy/.

Gendron, Angela. 2005. "Just War, Just Intelligence: An Ethical Framework for Foreign Espionage." *International Journal of Intelligence and CounterIntelligence* 18 (3): 398–434. https://doi.org/10.1080/08850600590945399.

Gershgorn, Dave. 2017. "Inside the Mechanical Brain of the World's First Robot Citizen." *Quartz* (blog). 2017. https://qz.com/1121547/how-smart-is-the-first-robot-citizen/.

Gilpin, L. H., D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. 2018. "Explaining Explanations: An Overview of Interpretability of Machine Learning." In *2018 IEEE 5th International Conference on Data*

Science and Advanced Analytics (DSAA), 80–89. https://doi.org/10.1109/DSAA.2018.00018.

Gips, James. 1994. "Toward the Ethical Robot." In *Android Epistemology*, edited by Kenneth M. Ford, C. Glymour, and Patrick Hayes. MIT Press.

Greene, Joshua. 2013. *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them*. 1st edition. New York: Penguin Press.

Greenwald, Glenn. 2013a. "NSA Collecting Phone Records of Millions of Verizon Customers Daily." *The Guardian*, June 6, 2013, sec. US news. https://www.theguardian.com/world/2013/jun/06/nsa-phone-records-verizon-court-order.

———. 2013b. "XKeyscore: NSA Tool Collects 'Nearly Everything a User Does on the Internet.'" *The Guardian*, July 31, 2013, sec. US news. https://www.theguardian.com/world/2013/jul/31/nsa-top-secret-program-online-data.

———. 2014. *No Place to Hide: Edward Snowden, the NSA, and the U.S. Surveillance State*. New York, NY: Metropolitan Books.

Greenwald, Glenn, and Ewen MacAskill. 2013. "NSA Prism Program Taps in to User Data of Apple, Google and Others." *The Guardian*, June 7, 2013, sec. US news. https://www.theguardian.com/world/2013/jun/06/us-tech-giants-nsa-data.

Gunkel, David J. 2014. "A Vindication of the Rights of Machines." *Philosophy & Technology* 27 (1): 113–32. https://doi.org/10.1007/s13347-013-0121-z.

Gutwirth, Serge, and Paul de Hert. 2006. "Privacy, Data Protection and Law Enforcement: Opacity of the Individual and Transparency of Power." In , edited by Eric Claes, Anthony Duff, and Serge Gutwirth, 61–104. Antwerp: Intersentia.

Haidt, Jonathan. 2001. "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment." *Psychological Review* 108 (4): 814–34. https://doi.org/10.1037/0033-295X.108.4.814.

Haidt, Jonathan, and Craig Joseph. 2008. "The Moral Mind: How Five Sets of Innate Intuitions Guide the Development of Many Culture-Specific Virtues, and Perhaps Even Modules." In *The Innate Mind: Volume 3: Foundations and the Future (Evolution and Cognition)*, edited by P. Carruthers, S. Laurence, and S. Stich. USA: Oxford University Press.

Harding, Luke. 2014. *The Snowden Files: The Inside Story Of The World's Most Wanted Man*. New York: Vintage Books.

Hardwig, John. 1991. "The Role of Trust in Knowledge." *The Journal of Philosophy* 88 (12): 693–708. https://doi.org/10.2307/2027007.

Hatmaker, Taylor. 2017. "Saudi Arabia Bestows Citizenship on a Robot Named Sophia." *TechCrunch* (blog). 2017. http://social.techcrunch.com/2017/10/26/saudi-arabia-robot-citizen-sophia/.

Heikoop, Daniël D., Marjan Hagenzieker, Giulio Mecacci, Simeon Calvert, Filippo Santoni De Sio, and Bart van Arem. 2019. "Human Behaviour with Automated Driving Systems: A Quantitative Framework for Meaningful Human Control." *Theoretical Issues in Ergonomics Science* 20 (6): 711–30. https://doi.org/10.1080/1463922X.2019.1574931.

Helbing, Dirk, Bruno S. Frey, Gerd Gigerenzer, Ernst Hafen, Michael Hagner, Yvonne Hofstetter, Jeroen van den Hoven, Roberto V. Zicari, and Andrej Zwitter. 2017. "Will Democracy Survive Big Data and Artificial Intelligence?" February 2017. https://www.scientificamerican.com/article/will-democracy-survive-big-data-and-artificial-intelligence/.

Heller, Joseph. 2011. *Catch-22*. Random House.

Hemberg, Erik, Jacob Rosen, Geoff Warner, Sanith Wijesinghe, and Una-May O'Reilly. 2016. "Detecting Tax Evasion: A Co-Evolutionary Approach." *Artificial Intelligence and Law* 24 (2): 149–82. https://doi.org/10.1007/s10506-016-9181-6.

Henschke, Adam. 2017. *Ethics in an Age of Surveillance: Personal Information and Virtual Identities*. New York: Cambridge University Press.

———. 2019. "Information Technologies and Construction of Perpetrator Identities." In *The Routledge International Handbook of Perpetrator Studies*, edited by Susanne C. Knittel and Zachary J. Goldberg, 217–27. New York, NY: Routledge.

High Level Expert Group on AI. 2019. "Ethics Guidelines for Trustworthy AI." European Commission. April 8, 2019. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

Hill, Kashmir. 2020. "The Secretive Company That Might End Privacy as We Know It." *The New York Times*, January 18, 2020, sec. Technology. https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html.

Himma, Kenneth Einar. 2009. "Artificial Agency, Consciousness, and the Criteria for Moral Agency: What Properties Must an Artificial Agent Have to Be

a Moral Agent?" *Ethics and Information Technology* 11 (1): 19–29. https://doi.org/10.1007/s10676-008-9167-5.

Horowitz, Michael C., and Paul Scharre. 2015. "Meaningful Human Control in Weapons Systems: A Primer." Center for a New American Security. 2015. https://s3.amazonaws.com/files.cnas.org/documents/Ethical_Autonomy_Working_Paper_031315.pdf?mtime=20160906082316.

Horwitz, Josh. 2017. "Facebook Says It Can Stop the Sharing of Most Terror-Related Posts within an Hour of Creation." Quartz. November 29, 2017. https://qz.com/1140539/facebook-says-its-able-to-stop-the-sharing-of-most-isis-terror-posts-within-an-hour-of-creation/.

Hoven, Jeroen van den. 2007. "ICT And Value Sensitive Design." In *The Information Society: Innovation, Legitimacy, Ethics And Democracy In Honor Of Professor Jacques Berleur s.j.*, edited by Philippe Goujon, Sylvian Lavelle, Penny Duquenoy, and Kai Kimppa, 233:67–72. Boston: Springer.

Hurka, Thomas. 2005. "Proportionality in the Morality of War." *Philosophy & Public Affairs* 33 (1): 34–66. https://doi.org/10.1111/j.1088-4963.2005.00024.x.

Interpol. 2017. "Speaker Identification Integrated Project (SIIP)." July 2017. https://www.interpol.int/en/Who-we-are/Legal-framework/Information-communications-and-technology-ICT-law-projects/Speaker-Identification-Integrated-Project-SIIP.

Joh, Elizabeth E. 2017. "Feeding the Machine: Policing, Crime Data, & Algorithms." *William & Mary Bill of Rights Journal* 26: 287.

Johnson, Deborah G. 2006a. "Computer Systems: Moral Entities but Not Moral Agents." *Ethics and Information Technology* 8 (4): 195–204. https://doi.org/10.1007/s10676-006-9111-5.

———. 2006b. "Computer Systems: Moral Entities but Not Moral Agents." *Ethics and Information Technology* 8 (4): 195–204. https://doi.org/10.1007/s10676-006-9111-5.

Johnson, Deborah G., and Keith W. Miller. 2008. "Un-Making Artificial Moral Agents." *Ethics and Information Technology* 10 (2): 123–33. https://doi.org/10.1007/s10676-008-9174-6.

Johnson, Justin M., and Taghi M. Khoshgoftaar. 2019. "Survey on Deep Learning with Class Imbalance."

Journal of Big Data 6 (1): 27. https://doi.org/10.1186/s40537-019-0192-5.

Jordan, M. I., and T. M. Mitchell. 2015. "Machine Learning: Trends, Perspectives, and Prospects." *Science* 349 (6245): 255–60. https://doi.org/10.1126/science.aaa8415.

Kasperkevic, Jana. 2015. "Google Says Sorry for Racist Auto-Tag in Photo App." *The Guardian*, July 1, 2015, sec. Technology. https://www.theguardian.com/technology/2015/jul/01/google-sorry-racist-auto-tag-photo-app.

Kendrick, Molly. 2019. "The Border Guards You Can't Win over with a Smile." BBC News. April 2019. https://www.bbc.com/future/article/20190416-the-ai-border-guards-you-cant-reason-with.

Klare, B. F., M. J. Burge, J. C. Klontz, R. W. Vorder Bruegge, and A. K. Jain. 2012. "Face Recognition Performance: Role of Demographic Information." *IEEE Transactions on Information Forensics and Security* 7 (6): 1789–1801. https://doi.org/10.1109/TIFS.2012.2214212.

Koepke, Logan. 2016. "Predictive Policing Isn't About the Future." Slate. November 21, 2016. http://www.slate.com/articles/technology/future_tense/2016/11/predictive_policing_is_too_dependent_on_historical_data.html.

Kofman, Ava. 2018. "Finding Your Voice: Forget About Siri and Alexa — When It Comes to Voice Identification, the 'NSA Reigns Supreme.'" *The Intercept* (blog). January 19, 2018. https://theintercept.com/2018/01/19/voice-recognition-technology-nsa/.

Kravets, David. 2013. "Declassified Documents Prove NSA Is Tapping the Internet." August 2013. https://www.wired.com/2013/08/nsa-tapping-internet/.

Kristjánsson, Kristján. 2007. *Aristotle, Emotions, and Education*. Hampshire; Burlington: Ashgate.

Kuang, Cliff. 2017. "Can A.I. Be Taught to Explain Itself?" *The New York Times*, November 21, 2017, sec. Magazine. https://www.nytimes.com/2017/11/21/magazine/can-ai-be-taught-to-explain-itself.html.

Lashbrook, Angela. 2018. "AI-Driven Dermatology Could Leave Dark-Skinned Patients Behind." The Atlantic. August 16, 2018. https://www.theatlantic.com/health/archive/2018/08/machine-learning-dermatology-skin-color/567619/.

Lee, Micah Lee, Glenn Greenwald, and Morgan Marquis-Boire. 2015. "A Look at the Inner Workings of NSA's XKEYSCORE." The Intercept. 2015. https://theintercept.com/2015/07/02/look-under-hood-xkeyscore/.

Levin, Sam. 2018. "Tesla Fatal Crash: 'autopilot' Mode Sped up Car before Driver Killed, Report Finds." The Guardian. June 7, 2018. https://www.theguardian.com/technology/2018/jun/07/tesla-fatal-crash-silicon-valley-autopilot-mode-report.

Litjens, Geert, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. 2017. "A Survey on Deep Learning in Medical Image Analysis." *Medical Image Analysis* 42 (December): 60–88. https://doi.org/10.1016/j.media.2017.07.005.

Lokhorst, Gert-Jan, and Jeroen van den Hoven. 2011. "Responsibility for Military Robots." In *Robot Ethics: The Ethical and Social Implications of Robotics*, edited by Patrick Lin, Keith Abney, and George A. Bekey, 145–56. MIT Press.

Macnish, Kevin. 2014. "Just Surveillance? Towards a Normative Theory of Surveillance." *Surveillance & Society* 12 (1): 142–53.

Markoff, John. 2015. "Relax, the Terminator Is Far Away." *The New York Times*, May 25, 2015. http://www.nytimes.com/2015/05/26/science/darpa-robotics-challenge-terminator.html.

Marquis-Boire, Morgan, Glenn Greenwald, and Micah Lee. 2015. "NSA's Google for the World's Private Communications." The Intercept. 2015. https://theintercept.com/2015/07/01/nsas-google-worlds-private-communications/.

Martin, Taylor, and David Priest. 2017. "The Complete List of Alexa Commands so Far." CNET. December 18, 2017. https://www.cnet.com/how-to/amazon-echo-the-complete-list-of-alexa-commands/.

McGoogan, Cara. 2016. "'You're Killing People': Elon Musk Attacks Critics of Self-Driving Cars." The Telegraph. October 20, 2016. https://www.telegraph.co.uk/technology/2016/10/20/youre-killing-people-elon-musk-attacks-critics-of-self-driving-c/.

Mecacci, Giulio, and Filippo Santoni de Sio. 2020. "Meaningful Human Control as Reason-Responsiveness: The Case of Dual-Mode Vehicles." *Ethics and*

Information Technology 22 (2): 103–15. https://doi.org/10.1007/s10676-019-09519-w.

Merritt, Maria. 2000. "Virtue Ethics and Situationist Personality Psychology." *Ethical Theory and Moral Practice* 3 (4): 365–83. https://doi.org/10.1023/A:1009926720584.

Metz, Cade. 2016. "In Two Moves, AlphaGo and Lee Sedol Redefined the Future." *Wired*, March 16, 2016. https://www.wired.com/2016/03/two-moves-alphago-lee-sedol-redefined-future/.

Miller, Keith W., Marty J. Wolf, and Frances Grodzinsky. 2017. "This 'Ethical Trap' Is for Roboticists, Not Robots: On the Issue of Artificial Agent Ethical Decision-Making." *Science and Engineering Ethics* 23 (2): 389–401. https://doi.org/10.1007/s11948-016-9785-y.

Miller, Seumas. 2008. *Terrorism and Counter-Terrorism: Ethics and Liberal Democracy*. Wiley. https://www.wiley.com/en-us/Terrorism+and+Counter+Terrorism%3A+Ethics+and+Liberal+Democracy-p-9781405139434.

Mo, H., X. Meng, J. Li, and S. Zhao. 2017. "Terrorist Event Prediction Based on Revealing Data." In *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)(*, 239–44. https://doi.org/10.1109/ICBDA.2017.8078815.

Moor, J. H. 2006. "The Nature, Importance, and Difficulty of Machine Ethics." *IEEE Intelligent Systems* 21 (4): 18–21. https://doi.org/10.1109/MIS.2006.80.

Moor, James. 2009. "Four Kinds of Ethical Robots." *Philosophy Now*, 2009.

Morrell, Alex. 2018. "Citigroup Has Inked a Deal with an AI-Powered Fintech to Help Flag Suspicious Payments and Safeguard a $4 Trillion Daily Operation." Business Insider. 2018. https://www.businessinsider.com/citi-has-inked-a-deal-with-an-ai-powered-fintech-feedzai-2018-12.

Mozur, Paul, Jonah M. Kessel, and Melissa Chan. 2019. "Made in China, Exported to the World: The Surveillance State." *The New York Times*, April 24, 2019, sec. Technology. https://www.nytimes.com/2019/04/24/technology/ecuador-surveillance-cameras-police-government.html.

Müller, Vincent C., and Nick Bostrom. 2016. "Future Progress in Artificial Intelligence: A Survey of Expert Opinion." In *Fundamental Issues of Artificial Intelligence*, edited by Vincent C. Müller, 555–72. Synthese Library 376. Springer International

Publishing. https://doi.org/10.1007/978-3-319-26485-1_33.
Nadella, Satya. 2016. "Microsoft's CEO Explores How Humans and A.I. Can Solve Society's Challenges—Together." Slate Magazine. June 28, 2016. https://slate.com/technology/2016/06/microsoft-ceo-satya-nadella-humans-and-a-i-can-work-together-to-solve-societys-challenges.html.
Nagel, Thomas. 2002. "Rawls and Liberalism." In , edited by Samuel Freeman, 62–85. Cambridge, U.K. ; New York: Cambridge University Press.
Nagenborg, Michael. n.d. "Artificial Moral Agents: An Intercultural Perspective," 6.
Nissenbaum, H. 2001. "How Computer Systems Embody Values." *Computer -Los Almalitos-* 34: 120.
Omand, David. 2010. *Securing the State*. London: C Hurst & Co Publishers Ltd.
Omand, Sir David, and Mark Phythian. 2013. "Ethics and Intelligence: A Debate." *International Journal of Intelligence and CounterIntelligence* 26 (1): 38–63. https://doi.org/10.1080/08850607.2012.705186.
Overly, Steven. 2017. "Facebook Plans to Use AI to Identify Terrorist Propaganda." *Washington Post*, February 16, 2017, sec. Innovations. https://www.washingtonpost.com/news/innovations/wp/2017/02/16/facebook-plans-to-use-ai-to-identify-terrorist-propaganda/.
Partnership on AI. 2019. "About." The Partnership on AI. 2019. https://www.partnershiponai.org/about/.
Pasquale, Frank. 2015. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge, Mass.: Harvard University Press. http://www.jstor.org/stable/j.ctt13x0hch.
Pieters, Janene. 2016. "Proposed Law Allows Massive Data Mining by Intelligence Agencies." NL Times. April 15, 2016. https://nltimes.nl/2016/04/15/proposed-law-allows-massive-data-mining-intelligence-agencies.
Pizarro, David. 2000. "Nothing More than Feelings? The Role of Emotions in Moral Judgment." *Journal for the Theory of Social Behaviour* 30 (4): 355–75. https://doi.org/10.1111/1468-5914.00135.
Poel, Ibo van de. 2013a. "Translating Values into Design Requirements." In *Philosophy and Engineering: Reflections on Practice, Principles, and Process*, edited by D. Mitchfelder, N. McCarty, and D.E. Goldberg. Dordrecht: Springer.

——. 2013b. “Translating Values into Design Requirements.” In *Philosophy and Engineering: Reflections on Practice, Principles, and Process*, edited by D. Mitchfelder, N. McCarty, and D.E. Goldberg, 253–66. Philosophy of Engineering and Technology 15. Dordrecht: Springer.

Presse, Agence France. 2018. “Computer Learns to Detect Skin Cancer More Accurately than Doctors.” *The Guardian*, May 29, 2018, sec. Society. https://www.theguardian.com/society/2018/may/29/skin-cancer-computer-learns-to-detect-skin-cancer-more-accurately-than-a-doctor.

Quinlan, Michael. 2007. “Just Intelligence: Prolegomena to an Ethical Theory.” *Intelligence and National Security* 22 (1): 1–13. https://doi.org/10.1080/02684520701200715.

Rainie, Lee, and Anderson, Janna. 2017. “Theme 7: The Need Grows for Algorithmic Literacy, Transparency and Oversight.”

Riedl, Mark O., and Brent Harrison. 2016. “Using Stories to Teach Human Values to Artificial Agents.” In *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*. https://www.aaai.org/ocs/index.php/WS/AAAIW16/paper/view/12624.

Robbins, Scott. 2019. “A Misdirected Principle with a Catch: Explicability for AI.” *Minds and Machines* 29 (4): 495–514. https://doi.org/10.1007/s11023-019-09509-3.

——. 2020. “AI and the Path to Envelopment: Knowledge as a First Step towards the Responsible Regulation and Use of AI-Powered Machines.” *AI & SOCIETY* 35 (2): 391–400. https://doi.org/10.1007/s00146-019-00891-1.

Robbins, Scott, and Adam Henschke. 2017. “The Value of Transparency: Bulk Data and Authoritarianism.” *Surveillance & Society* 15 (3/4): 582–89. https://doi.org/10.24908/ss.v15i3/4.6606.

Roberts, Dan. 2013. “FBI Chief Mueller Says Spy Tactics Could Have Stopped 9/11 Attacks.” *The Guardian*, December 2013.

Roeser, S. 2010. *Moral Emotions and Intuitions*. Springer.

Rolnick, David, Priya L. Donti, Lynn H. Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, et al. 2019. “Tackling Climate Change with Machine Learning.” *ArXiv:1906.05433 [Cs, Stat]*, November. http://arxiv.org/abs/1906.05433.

Roper, James E. 2010. "Using Private Corporations to Conduct Intelligence Activities for National Security Purposes: An Ethical Appraisal." *International Journal of Intelligence Ethics* 1 (2): 46–73.

Russell, Stuart J., and Peter Norvig. 1995. *Artificial Intelligence: A Modern Approach*. Prentice Hall Series in Artificial Intelligence. Englewood Cliffs, N.J: Prentice Hall.

Rutkin, Aviva. 2014. "Ethical Trap: Robot Paralysed by Choice of Who to Save." New Scientist. September 10, 2014. https://www.newscientist.com/article/mg22329863-700-ethical-trap-robot-paralysed-by-choice-of-who-to-save/.

Sampler, Ian. 2017. "Ban on Killer Robots Urgently Needed, Say Scientists." The Guardian. November 13, 2017. http://www.theguardian.com/science/2017/nov/13/ban-on-killer-robots-urgently-needed-say-scientists.

Santoni de Sio, Filippo, and Jeroen van den Hoven. 2018. "Meaningful Human Control over Autonomous Systems: A Philosophical Account." *Frontiers in Robotics and AI* 5. https://doi.org/10.3389/frobt.2018.00015.

Saunders, 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* 12 (3): 347–71. https://doi.org/10.1007/s11292-016-9272-0.

Savage, Charlie. 2017. "N.S.A. Warrantless Surveillance Aided Turks After Attack, Officials Say." *The New York Times*, June 2017.

Schauer, Frederick F. 2003. *Profiles, Probabilities, and Stereotypes*. Harvard University Press.

Scheutz, Matthias. 2016. "The Need for Moral Competency in Autonomous Agent Architectures." In *Fundamental Issues of Artificial Intelligence*, edited by Vincent C. Müller, 515–25. Synthese Library 376. Springer International Publishing. https://doi.org/10.1007/978-3-319-26485-1_30.

Schwartz, Mattathias. 2015. "The Whole Haystack." January 2015. http://www.newyorker.com/magazine/2015/01/26/whole-haystack.

Shafer-Landau, Russ. 1994. "Ethical Disagreement, Ethical Objectivism and Moral Indeterminacy." *Philosophy and Phenomenological Research* 54 (2): 331–44. https://doi.org/10.2307/2108492.

Sharkey, Amanda. 2016. "Should We Welcome Robot Teachers?" *Ethics and Information Technology* 18 (4): 283–97. https://doi.org/10.1007/s10676-016-9387-z.
———. 2017. "Can We Program or Train Robots to Be Good?" *Ethics and Information Technology*, May, 1–13. https://doi.org/10.1007/s10676-017-9425-5.
Sharkey, Noel. 2008. "The Ethical Frontiers of Robotics." *Science* 322 (5909): 1800–1801. https://doi.org/10.1126/science.1164582.
———. 2011. "Automating Warfare: Lessons Learned from the Drones." *Journal of Law, Information & Science* 21 (2). https://doi.org/10.5778/JLIS.2011.21.Sharkey.1.
———. 2012. "The Evitability of Autonomous Robot Warfare." *International Review of the Red Cross* 94 (886): 787–799. https://doi.org/10.1017/S1816383112000732.
———. 2014. "Towards a Principle for the Human Supervisory Control of Robot Weapons." *Politica & Società*, no. 2/2014. https://doi.org/10.4476/77105.
Sharkey, Noel, and Amanda Sharkey. 2011. "The Rights and Wrongs of Robot Care." In *Robot Ethics: The Ethical and Social Implications of Robotics*, edited by Patrick Lin, Keith Abney, and George A. Bekey, 267–82. MIT Press.
Sharkey, Noel, Aimee van Wynsberghe, Scott Robbins, and Eleanor Hancock. 2017. "Our Sexual Future with Robots." Foundationn for Responsible Robotics. July 5, 2017. https://responsiblerobotics.org/wp-content/uploads/2017/07/FRR-Consultation-Report-Our-Sexual-Future-with-robots_Final.pdf.
Shirky, Clay. 2009. "A Speculative Post on the Idea of Algorithmic Authority." *Clay Shirky* (blog). November 15, 2009. http://www.shirky.com/weblog/2009/11/a-speculative-post-on-the-idea-of-algorithmic-authority/.
Silver, David, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, et al. 2016. "Mastering the Game of Go with Deep Neural Networks and Tree Search." *Nature* 529 (7587): 484–89. https://doi.org/10.1038/nature16961.
Simon, Judith. 2010. "The Entanglement of Trust and Knowledge on the Web." *Ethics and Information Technology* 12 (4): 343–55. https://doi.org/10.1007/s10676-010-9243-5.
Skitka, LINDA J., KATHLEEN L. Mosier, and MARK Burdick. 1999. "Does Automation Bias Decision-Making?"

International Journal of Human-Computer Studies 51 (5): 991–1006. https://doi.org/10.1006/ijhc.1999.0252.

Solove, Daniel J. 2004. *The Digital Person: Technology And Privacy In The Information Age*. First Edition edition. New York: NYU Press.

Street, Sharon. 2006. "A Darwinian Dilemma for Realist Theories of Value." *Philosophical Studies* 127 (1): 109–66. https://doi.org/10.1007/s11098-005-1726-6.

Sulleyman, Aatif. 2018. "Boston Dynamics Robot Dog Opens Door for Another Robot with No Arms." The Independent. February 13, 2018. http://www.independent.co.uk/life-style/gadgets-and-tech/news/boston-dynamics-robot-dog-open-door-video-hold-open-no-arms-help-video-youtube-a8208096.html.

Sweet, Matthew. 2020. "Rewind - How White Nationalists Hijacked the Hawaiian Shirt | 1843." The Economist. 2020. https://www.economist.com/1843/2020/07/31/how-white-nationalists-hijacked-the-hawaiian-shirt.

The Economist. 2012. "Morals and the Machine." The Economist. 2012. https://www.economist.com/leaders/2012/06/02/morals-and-the-machine.

The New Yorker. 2014. *The Virtual Interview: Edward Snowden - The New Yorker Festival*. https://www.youtube.com/watch?v=fidq3jow8bc.

The Public Voice. 2018. "AI Universal Guidelines – Thepublicvoice.Org." 2018. https://thepublicvoice.org/ai-universal-guidelines/.

Thomas, Elise. 2019. "How to Hack Your Face to Dodge the Rise of Facial Recognition Tech." *Wired UK*, February 1, 2019. https://www.wired.co.uk/article/avoid-facial-recognition-software.

Thompson, Avery. 2019. "Five AIs Just Worked Together To Beat a Top Human Video Game Team." Popular Mechanics. April 16, 2019. https://www.popularmechanics.com/technology/robots/a27156719/openai-dota-2-victory/.

Tonkens, Ryan. 2009. "A Challenge for Machine Ethics." *Minds and Machines* 19 (3): 421. https://doi.org/10.1007/s11023-009-9159-1.

Travis, Alan. 2016. "'Snooper's Charter' Bill Becomes Law, Extending UK State Surveillance." The Guardian. November 29, 2016. http://www.theguardian.com/world/2016/nov/29/snoop

ers-charter-bill-becomes-law-extending-uk-state-surveillance.
Turek, Matt. n.d. "Explainable Artificial Intelligence." Defense Advanced Research Projects Agency. Accessed August 25, 2020. https://www.darpa.mil/program/explainable-artificial-intelligence.
Turilli, Matteo, and Luciano Floridi. 2009. "The Ethics of Information Transparency." *Ethics and Information Technology* 11 (2): 105–12. https://doi.org/10.1007/s10676-009-9187-9.
Turing, Alan M. 1950. "Computing Machinery and Intelligence." *Mind* 59 (236): 433–460.
Uddin, M. Irfan, Nazir Zada, Furqan Aziz, Yousaf Saeed, Asim Zeb, Syed Atif Ali Shah, Mahmoud Ahmad Al-Khasawneh, and Marwan Mahmoud. 2020. "Prediction of Future Terrorist Activities Using Deep Neural Networks." Research Article. Complexity. Hindawi. April 22, 2020. https://doi.org/10.1155/2020/1373087.
UNI Global Union. 2018. "10 Principles for Ethical AI." UNI Global Union. http://www.thefutureworldofwork.org/opinions/10-principles-for-ethical-ai/.
United States Department of Defense. 2012. "Department of Defense Directive on Autonomouse Weapons Systems." https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodd/300009p.pdf.
US State Department. 2018. "Country Reports on Terrorism 2017." US State Department. https://www.state.gov/reports/country-reports-on-terrorism-2017/.
Ustun, Berk, Alexander Spangher, and Yang Liu. 2019. "Actionable Recourse in Linear Classification." In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 10–19. FAT* '19. New York, NY, USA: ACM. https://doi.org/10.1145/3287560.3287566.
Vallor, Shannon. 2015. "Moral Deskilling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character." *Philosophy & Technology* 28 (1): 107–24. https://doi.org/10.1007/s13347-014-0156-9.
Vincent, James. 2018. "AI That Detects Cardiac Arrests during Emergency Calls Will Be Tested across Europe This Summer." The Verge. April 25, 2018. https://www.theverge.com/2018/4/25/17278994/ai-cardiac-arrest-corti-emergency-call-response.

——. 2019. “This Colorful Printed Patch Makes You Pretty Much Invisible to AI.” The Verge. April 23, 2019. https://www.theverge.com/2019/4/23/18512472/fool-ai-surveillance-adversarial-example-yolov2-person-detection.

Vogelstein, Fred, and Will Knight. 2020. “Health Officials Say ‘No Thanks’ to Contact-Tracing Tech.” *Wired*, May 8, 2020. https://www.wired.com/story/health-officials-no-thanks-contact-tracing-tech/.

Wachter, Sandra, Brent Mittelstadt, and Luciano Floridi. 2016. “Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation.” *International Data Privacy Law* 7 (2): 76–99. https://doi.org/10.1093/idpl/ipx005.

Wachter, Sandra, Brent Mittelstadt, and Chris Russell. 2017. “Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR.” *Harvard Journal of Law & Technology* 31 (2). http://arxiv.org/abs/1711.00399.

Waldrop, M. Mitchell. 1987. “A Question of Responsibility.” *AI Magazine* 8 (1): 28–28. https://doi.org/10.1609/aimag.v8i1.572.

Wallach, Wendall, and Colin Allen. 2010. *Moral Machines: Teaching Robots Right from Wrong*. 1 edition. New York: Oxford University Press.

Wallach, Wendell. 2007. “Implementing Moral Decision Making Faculties in Computers and Robots.” *AI & SOCIETY* 22 (4): 463–75. https://doi.org/10.1007/s00146-007-0093-6.

——. 2010. “Robot Minds and Human Ethics: The Need for a Comprehensive Model of Moral Decision Making.” *Ethics and Information Technology* 12 (3): 243–50. https://doi.org/10.1007/s10676-010-9232-8.

West, Leah. 2018. “Canada Tries Domestic Bulk Collection: It Just Might Work.” Lawfare. March 26, 2018. https://lawfareblog.com/canada-tries-domestic-bulk-collection-it-just-might-work.

Whittaker, Meredeth, Kate Crawford, Roel Dobbe, Genevieve Fried, Elizabeth Kaziunas, Varoon Mathur, Sarah Myers West, Rashida Richardson, Jason Shultz, and Oscar Schwartz. 2018. “AI Now 2018.” AI Now Institute. December 2018. https://ainowinstitute.org/AI_Now_2018_Report.html.

Wiegel, Vincent. 2006. “Building Blocks for Artificial Moral Agents.” *Proceedings of EthicalALife06 Workshop*, January.

——. 2010. “Wendell Wallach and Colin Allen: Moral Machines: Teaching Robots Right from Wrong.” *Ethics and Information Technology* 12 (4): 359–61. https://doi.org/10.1007/s10676-010-9239-1.

Winter, Joost C. F. de, Riender Happee, Marieke H. Martens, and Neville A. Stanton. 2014. “Effects of Adaptive Cruise Control and Highly Automated Driving on Workload and Situation Awareness: A Review of the Empirical Evidence.” *Transportation Research Part F: Traffic Psychology and Behaviour*, Vehicle Automation and Driver Behaviour, 27 (November): 196–217. https://doi.org/10.1016/j.trf.2014.06.016.

Wynsberghe, Aimee van. 2012. “Designing Robots with Care: Creating an Ethical Framework for the Future Design and Implementation of Care Robots.”

——. 2013. “Designing Robots for Care: Care Centered Value-Sensitive Design.” *Science and Engineering Ethics* 19 (2): 407–33. https://doi.org/10.1007/s11948-011-9343-6.

——. 2016a. *Healthcare Robots: Ethics, Design and Implementation*. New York, NY: Routledge.

——. 2016b. “Service Robots, Care Ethics, and Design.” *Ethics and Information Technology* 18 (4): 311–21. https://doi.org/10.1007/s10676-016-9409-x.

Wynsberghe, Aimee van, and Scott Robbins. 2014. “Ethicist as Designer: A Pragmatic Approach to Ethics in the Lab.” *Science and Engineering Ethics* 20 (4): 947–61. https://doi.org/10.1007/s11948-013-9498-4.

——. 2019. “Critiquing the Reasons for Making Artificial Moral Agents.” *Science and Engineering Ethics* 25 (3): 719–35. https://doi.org/10.1007/s11948-018-0030-8.

Acknowledgements

I'd like to thank the many people who helped make this dissertation possible.

First, I would like to thank my supervisor Seumas. You placed a lot of trust in me that I would get this done. You allowed me to be, at times, extremely independent. And when the time came, you were available for all the feedback and criticism I needed. Thank you.

To Paul, Adam, Mitt, Michael, Michael, Do'aa, Jonas, Alastair, and Tony – I have learned a lot from you all. The lively Q&A sessions after my talks were invaluable to the writing that ended up in this thesis. What a blast we had during all those meetings in Oxford, Georgetown, and The Hague.

Thanks to the people who inspired me (whether they know it or not) to take the path that I have. John Mahoney and Renee Renner from CSU Chico were the two people who showed me that there was an intersection between science, technology, and philosophy. Johnny and Lynn – you made me believe that I could achieve something like this. Jeroen vdH, Ibo vdP and Mark Alfano – thank you for the encouragement, criticism, and support.

Noel and Amanda – you are my academic parents and two of my best friends. You have given so much to me and my family. I can't imagine finishing this PhD without you.

To the many friends that have helped me in countless ways during this process. Jacob, Kees, Louise, Taylor, Laura, Duuk, Bartell, Tina, Connor, Madelaine, David, Joanna, Lindy, Mark, Emma, Robert and many others. What great times we have all had. To my paranymphs Taylor and Duuk – you have become such great friends. Your help is the only reason that I will get this thesis printed in time.

To my extended family in the US and Canada for supporting me in the many ways only family can. To my dad – you instilled in me a love of argument. Without that, I could not have finished this. To my mom – you instilled in me a love of cooking – and provided many tools to aid my efforts. Which in the Netherlands I would have starved without! To my brother Bryan, you have become a great friend – and someone who reminds me of the great things that exist outside of academia. All of your continued reliance on every new gadget tells me I have lots more work to do!

To my children Aria and Holden – thanks for keeping my life in perspective. All the world's problems seem a lot less significant when you two are laughing.

Most importantly, I would like to thank my wife Aimee. None of this exists without you. My arguments in this thesis were developed during countless nights of having drinks and debating about AI and ethics on our patio. No matter how hard this process was – you made it fun. I love you.

Bonn, Germany, 26 December 2020

About the Author

Scott Alan Robbins (1984) was born in San Diego, California, USA. In February 2021 he will be a postdoctoral researcher affiliated with the Center for Advanced Security, Strategic and Integration Studies (CASSIS) at the University of Bonn in Germany. He completed his PhD in Ethics of Technology at Delft University of Technology between January 2017 and January 2021. Scott has an interdisciplinary academic and professional background. He holds a Bachelor of Science (Cum Laude) from California State University, Chico (2006), where he studied computer science. Before pursuing a master's degree, he worked for companies large and small as a computer scientist. Scott received a Master's of Science in Ethics and Technology from Twente University in the Netherlands (2013). Scott lives in Bonn, Germany with his wife and two children.

List of Publications

Robbins, S. "Bulk Data Collection Ethics" (forthcoming). In Miller, S. (ed), *Counter Terrorism: The Ethical Issues*. Edward Elgar. UK.

Robbins, S. "Machine Learning, National Security, and Ethics" (forthcoming).In Clarke, M., Henschke, A., & Legrand, T. (eds), *Palgrave Handbook of National Security*. Palgrave.

Robbins, S. "A Misdirected Principle with a Catch: Explicability for AI". (2019). *Minds and Machines*. https://doi.org/10.1007/s11023-019-09509-3 .

Robbins, S. "AI and the path to envelopment: knowledge as a first step towards the responsible regulation and use of AI-powered machines". (2019) *AI & SOCIETY*. https://doi.org/10.1007/s00146-019-00891-1.

Van Wynsberghe, A. & Robbins, S. "Critiquing the Reasons for Artificial Moral Agents", *Science and Engineering Ethics.* (2018) DOI https://doi.org/10.1007/s11948-018-0030-8.

Fosch Villaronga et. Al. & Robbins, S. "Nothing Comes between My Robot and Me': Privacy and Human-Robot Interaction in Robotised Healthcare" (2018) in Leenes, R. et al. (Eds.). *Data Protection and Privacy: The Internet of Bodies.*

Robbins, S. & Henschke, A. "Designing for Democracy: Bulk Data and Authoritarianism", *Surveillance and Society 15 (3/4)*. (2017): 582-589. DOI: https://doi.org/10.24908/ss.v15i3/4.6606.

Sharkey, N., van Wynsberghe, A., Robbins, S. & Hancock, E. "Our sexual future with robots". *Foundation for Responsible Robotics*. (2017). https://responsible-robotics-myxf6pn3xr.netdna-ssl.com/wp-content/uploads/2017/11/FRR-Consultation-Report-Our-Sexual-Future-with-robots-.pdf.

Van Wynsberghe, A. & Robbins, S. "Ethicist as Designer: a pragmatic approach to ethics in the lab," *Science and Engineering Ethics* 20 (4) (2014): 947-61. DOI: https://doi.org/10.1007/s11948-013-9498-4.